

Unfolding Spatial Cognition

Evaluating Multimodal Models on Visual Simulations

Linjie (Lindsey) Li
04/06/2026

Question

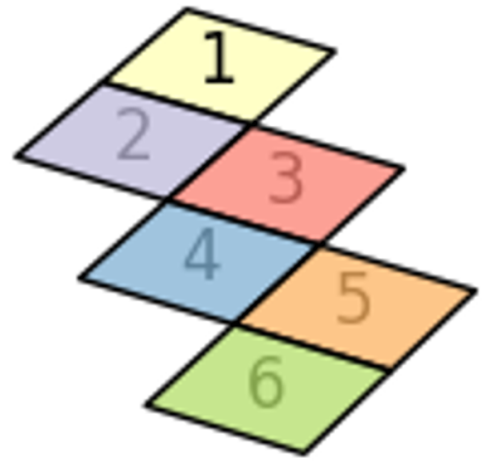


Can the net be folded to form a cube, yes or no?

Yes!

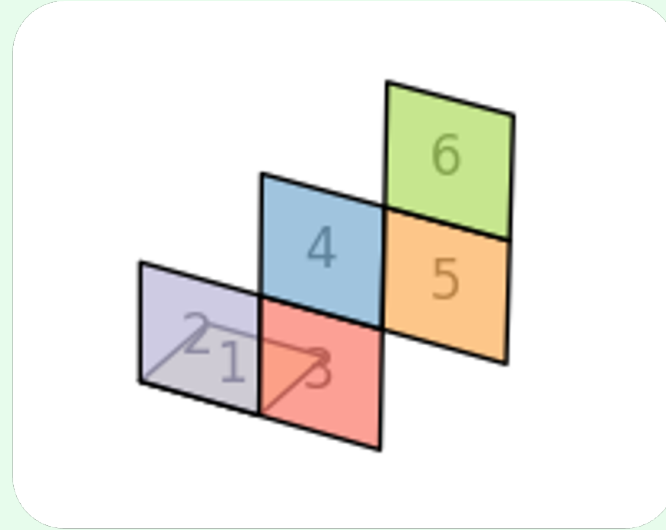


Question

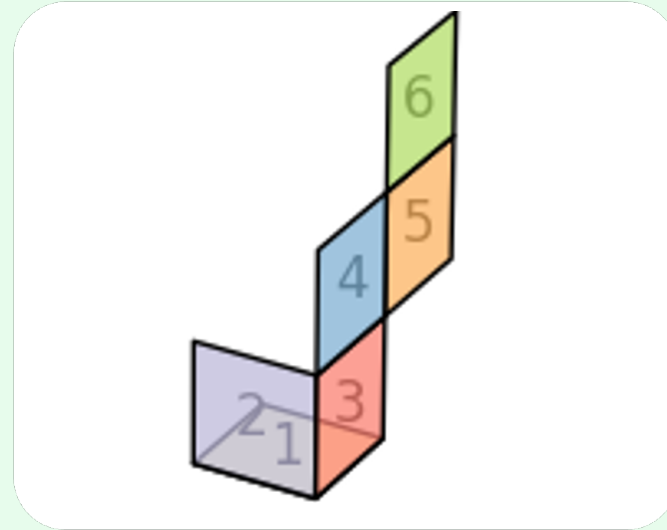


Can the net be folded to form a cube, yes or no?

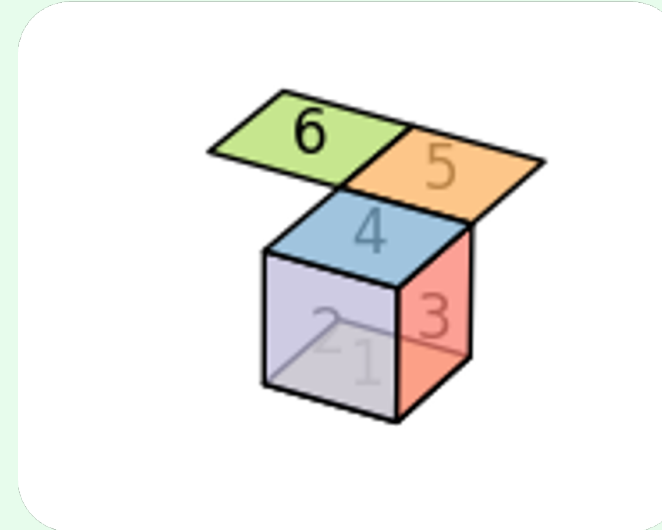
Step-by-Step Human Mental Simulation



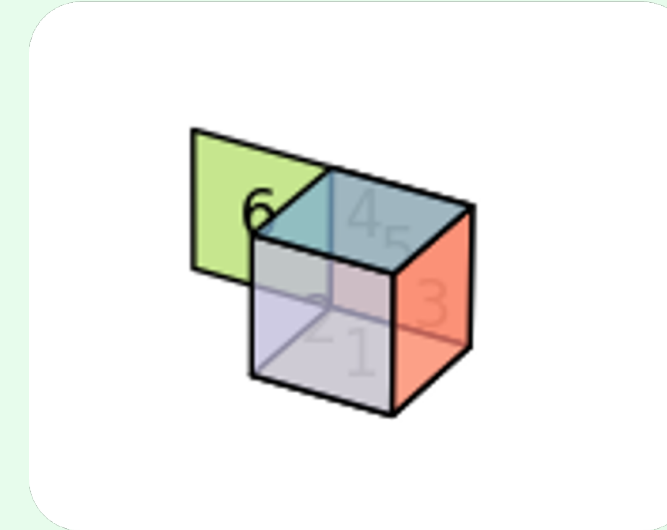
1



2



3



4



Yes!

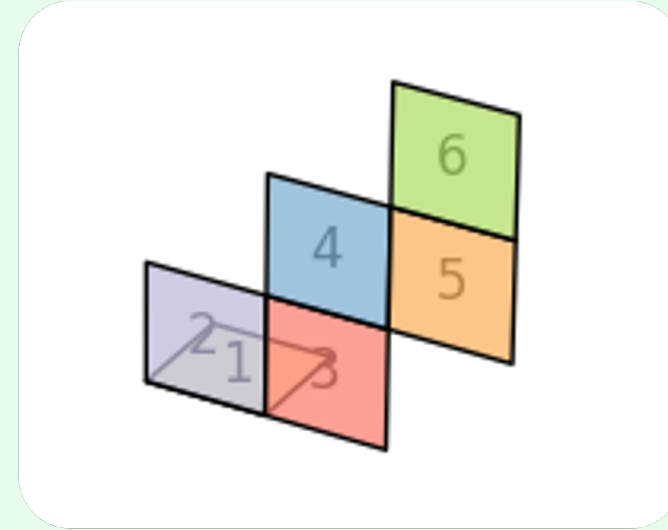


Question

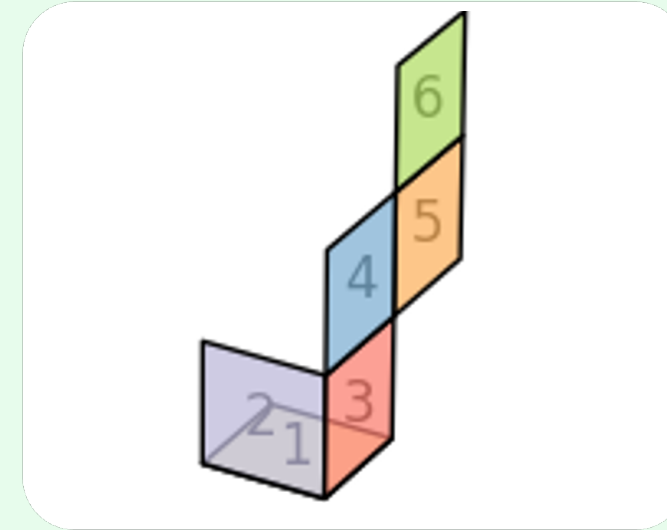


Can the net be folded to form a cube, yes or no?

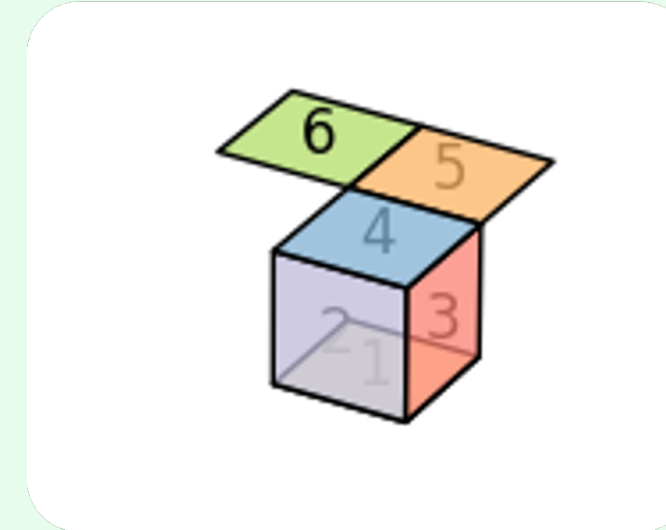
Step-by-Step Human Mental Simulation



1



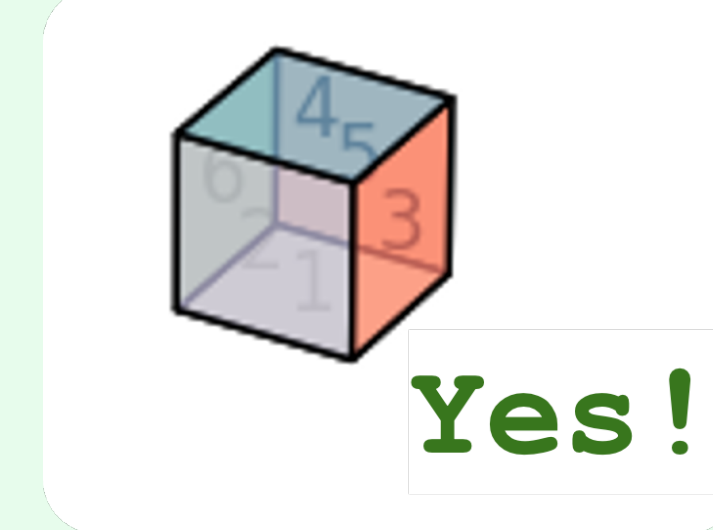
2



3



4



Yes!



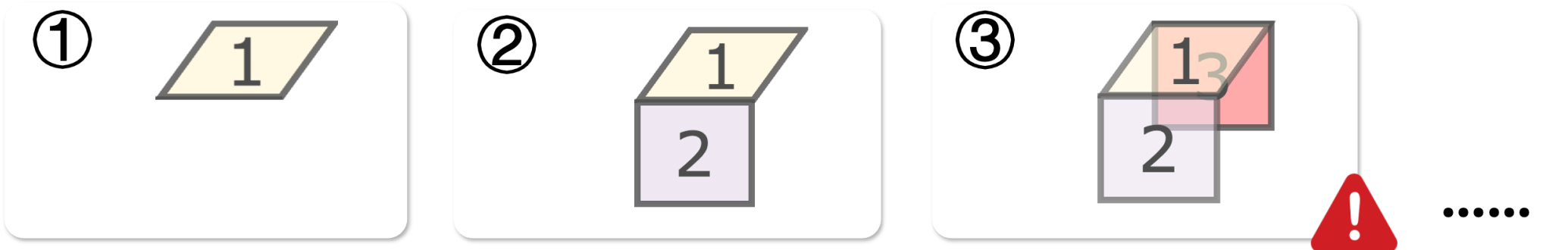
MLLM Textual Simulation

GPT-4o: Yes, this net can be folded into a cube. Here's how:

- ① Face 1 becomes the top of the cube.
- ② Face 2 folds down to the left of Face 1.
- ③ Face 3 folds down to the right of Face 1. } !
- ④ Face 4 folds down below Face 2, becoming the front face.
- ⑤ Face 5 folds down below Face 3, becoming the back face.
- ⑥ Face 6 folds up to close the bottom of the cube.

All faces will perfectly align ✓, forming a six-faced cube with no overlaps or gaps.

Visualization of Textual Reasoning:



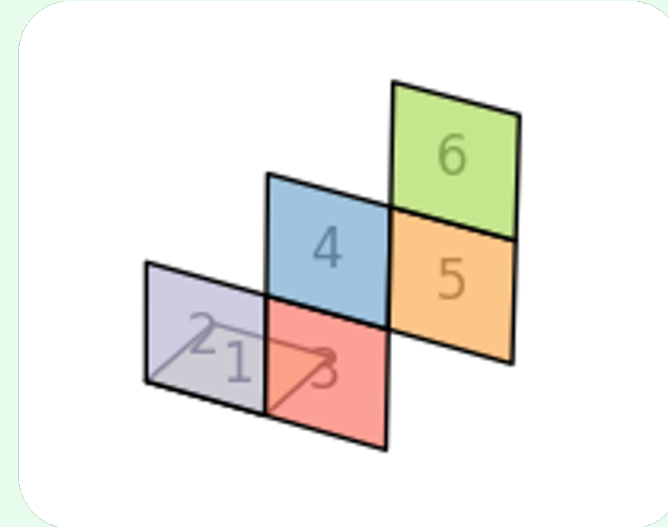
Reasoning Error Analysis: If Face 2 is on the **left** of Face 1 (②) and Face 3 is on the **right** (③), they would be opposite, but they are adjacent in the net. The correct third step should follow the net's layout to maintain their connection.

Question

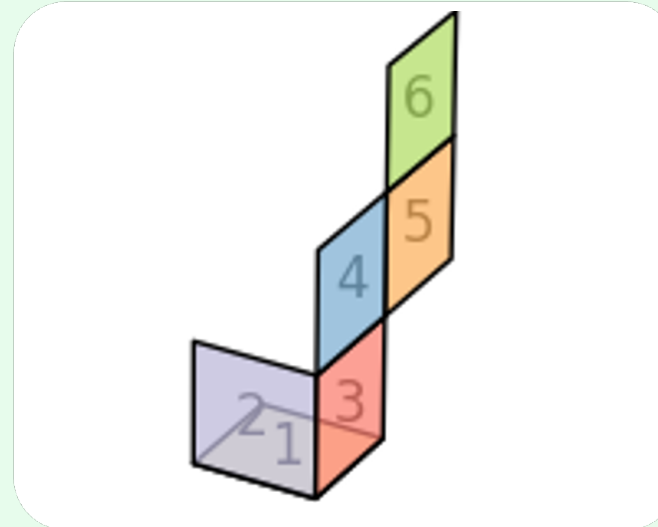


Can the net be folded to form a cube, yes or no?

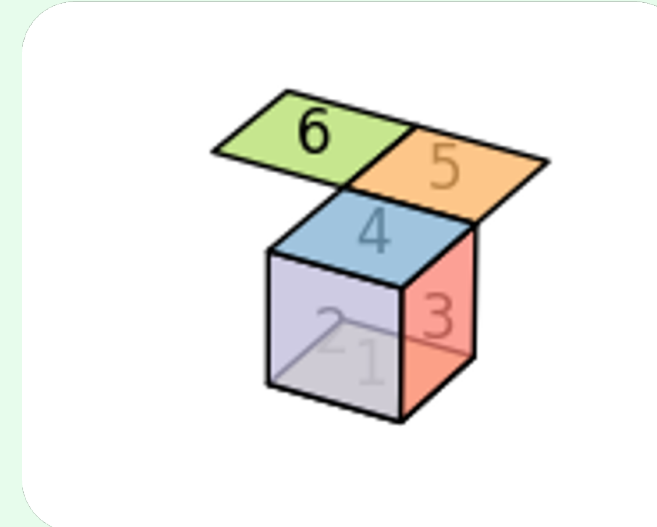
Step-by-Step Human Mental Simulation



1



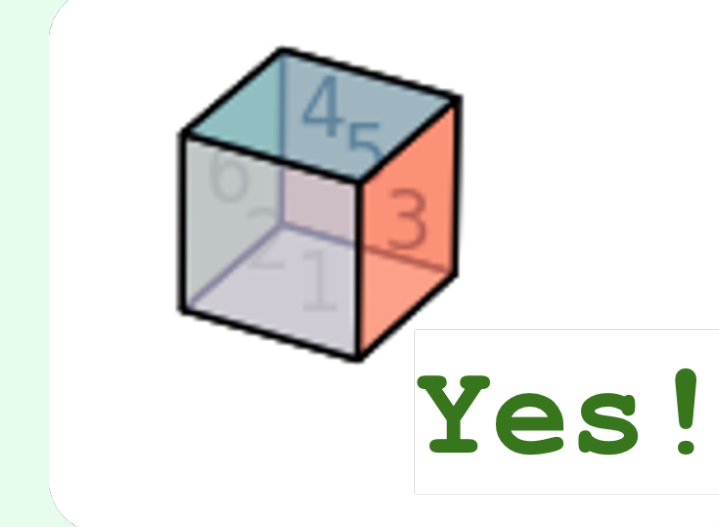
2



3



4



Yes!



MLLM Textual Simulation

O₃ (Thinking with Image):

...
There are 11 possible valid cube nets, and this one is shaped in a zigzag
...

Yes — this zig-zag arrangement can indeed be folded into a cube.

- ① Squares 2-3-4-5 wrap around to make the lateral “belt,”
- ② Square 1 folds up to become the top face
- ③ Square 6 folds under to become the bottom face.
- ④ Once the belt is closed (2 joins 5),



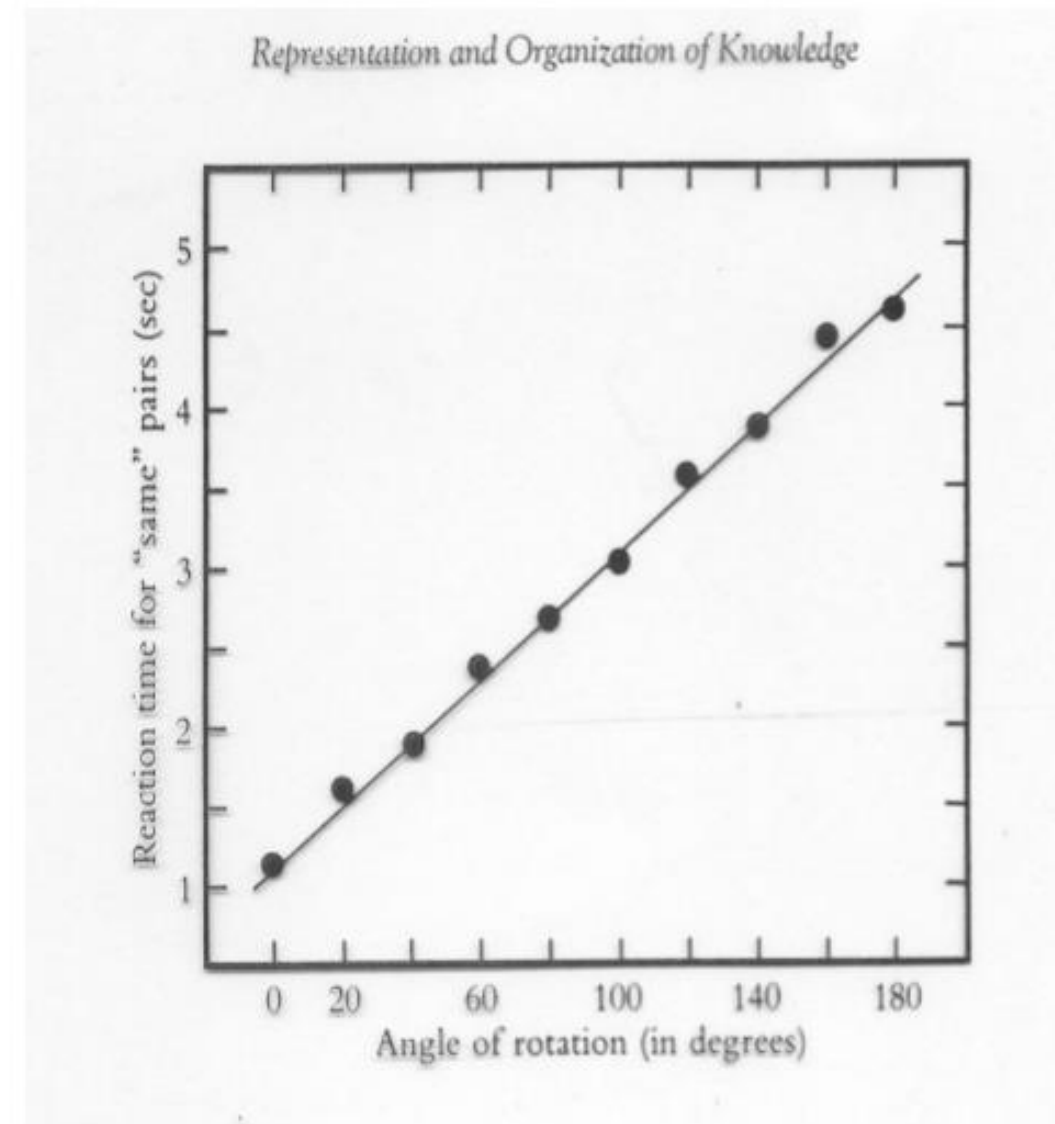
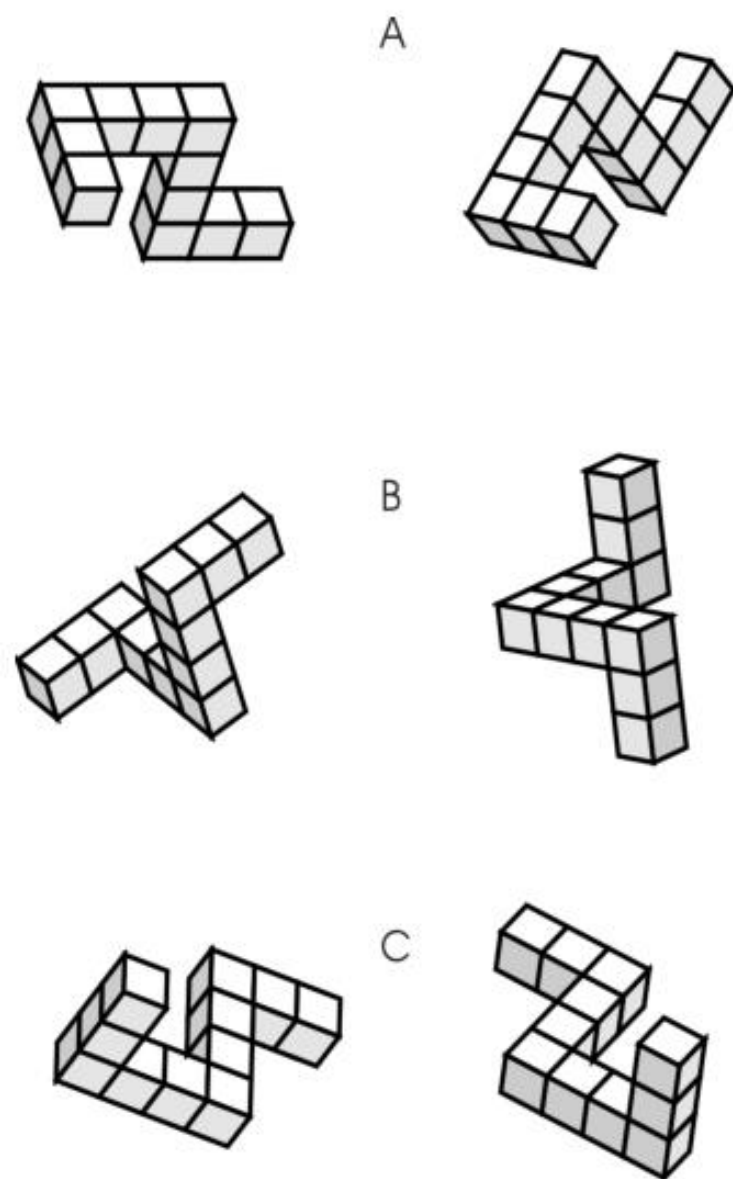
None of the faces overlap, ✓ satisfying the requirements for a valid cube net.

Reasoning Error Analysis:

- ② - ③ contradict with the fact that **Face 1 is adjacent to Face 6.**
- ④ contradict with the fact that **Face 2 is not adjacent to Face 5**

Visual Simulation is Critical to Human (non-verbal) Reasoning

Mental Rotation (Shepard & Metzler, 1971)



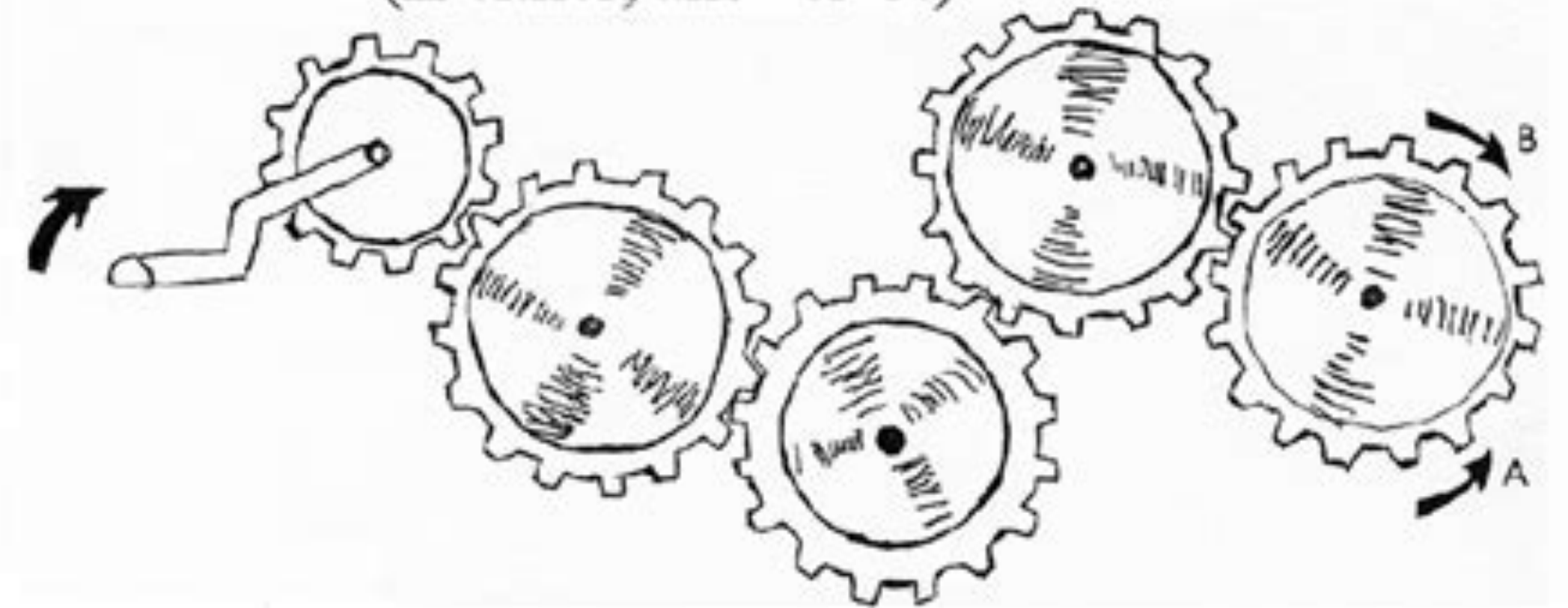
Claudia J. Stanny

9

Mechanical reasoning by mental simulation (Hegarty, 2004)

When the handle is turned in the direction shown, which direction will the final gear turn?

(If either, answer C.)



"Spatial ability predicts performance in mathematics and eventual expertise in science, technology and engineering." (Tosto, M. G. et al. 2014)



People with spatial intelligence - "skillfully use the ability to create images, spatial relationships, and visualizations in the mind." (Pawlak-Jakubowska & Terczyńska 2023)

Yet, Existing Benchmarks Focus on Recognition or Verbal Reasoning



Static Recognition

Perception (Coarse-Grained Tasks)



Existence 🤖

	[Y] Is there a elephant in this image? [N] Is there a hair drier in this image?		[Y] Is there a refrigerator in this image? [N] Is there a donut in this image?
---	--	--	---


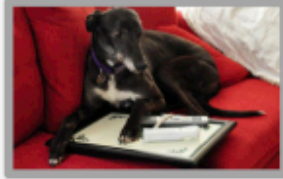
Count 🧮

	[Y] Is there a total of two person appear in the image? [N] Is there only one person appear in the image?		[Y] Are there two pieces of pizza in this image? [N] Is there only one piece of pizza in this image?
--	--	---	---

Position 📍

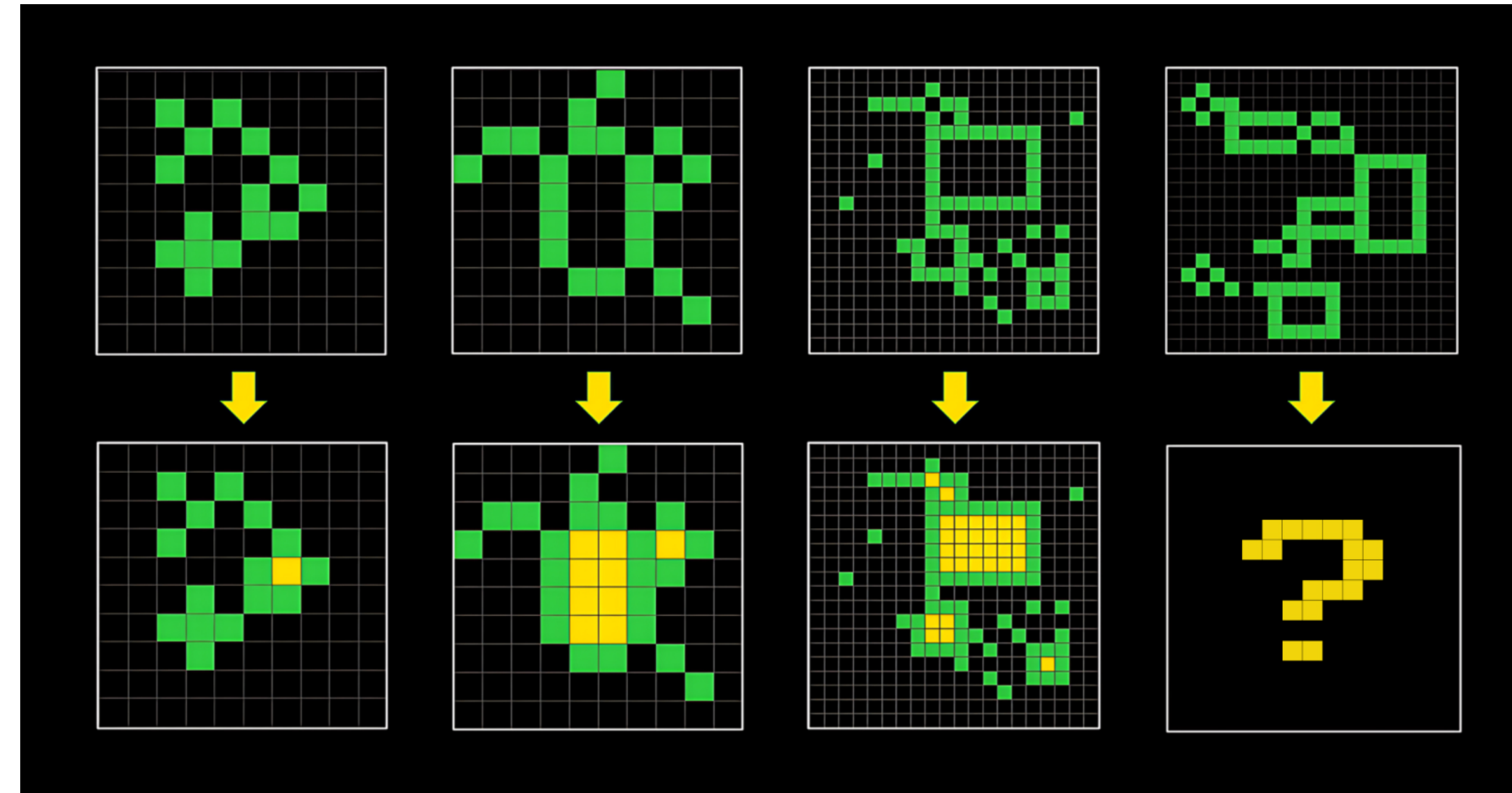
	[Y] Is the motorcycle on the right side of the bus? [N] Is the motorcycle on the left side of the bus.		[Y] Is the baby on the right of the dog in the image? [N] Is the baby on the left of the dog in the image?
---	---	--	---

Color 🌈

	[Y] Is there a red coat in the image? [N] Is there a yellow coat in the image?		[Y] Is there a red couch in the image? [N] Is there a black couch in the image?
---	---	--	--

MME
(Fu et al., 2023)

Problems that can be re-phrased to linguistic reasoning




ARC-AGI
(Chollet, 2019)

Yet, Existing Benchmarks Focus on Recognition or Verbal Reasoning

Relative depth

Which point is closer?


Relative reflectance

Which point is darker?

Functional correspondence

Which points have similar affordance when pulling out a nail?

Jigsaw

Which image fits here?

Multi-view reasoning

Is camera moving right?

Visual correspondence

Which point is the same?

Semantic correspondence

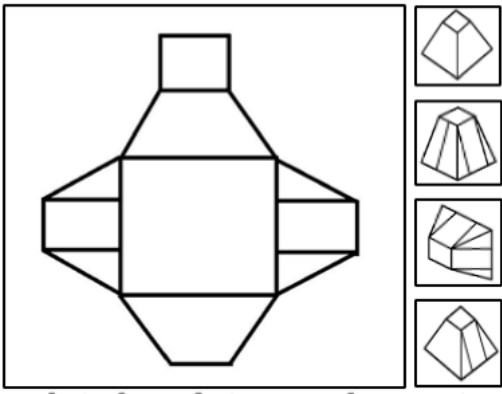
Which points have similar semantics?

Forensics detection

Which image is real?

Visual similarity

Which image is more similar to the left?

IQ Test

Which object does it fold into?

BLINK
Visual tasks beyond language descriptions

BLINK
(Fu et al., 2024)

STARE - Spatial Transformation And Reasoning Evaluation



Tasks that can be better solved using visual simulations (non-verbal reasoning).

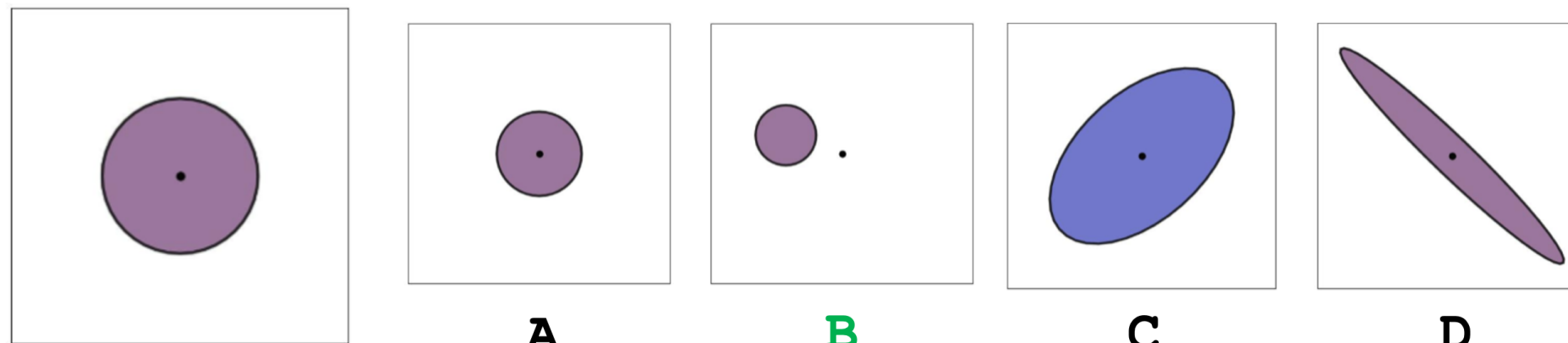
Tasks that requires multi-step visual reasoning rather than solving in a BLINK.

Tasks that are harder to be reduced to language reasoning - similar to BLINK.

STARE

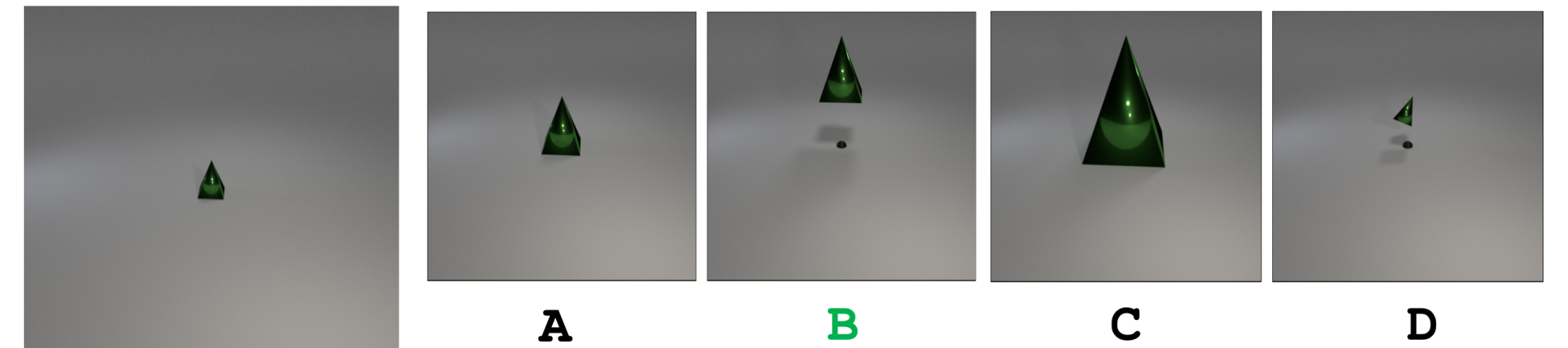
Foundational Geometric Transformations

2D Transformation



Make it smaller then move it left and slightly upwards.

3D Transformation

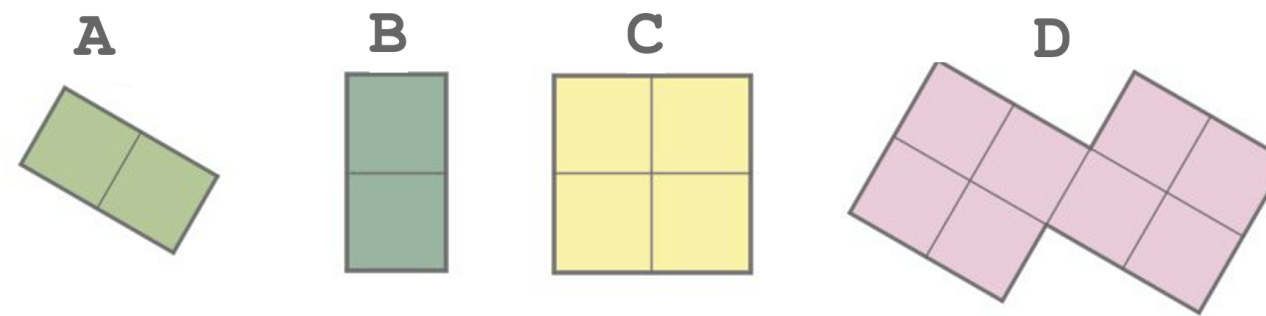
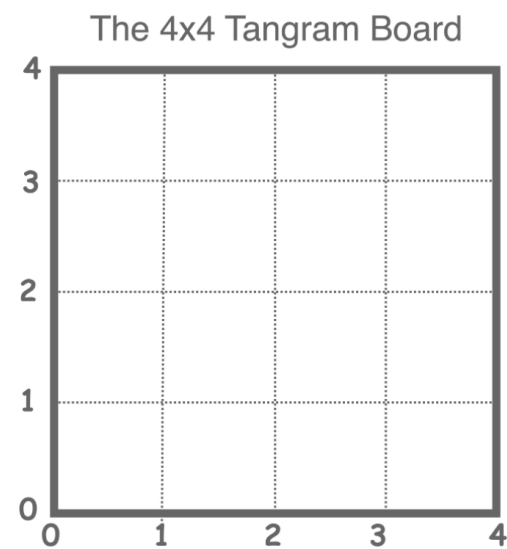


Move it up and then make it bigger.

STARE

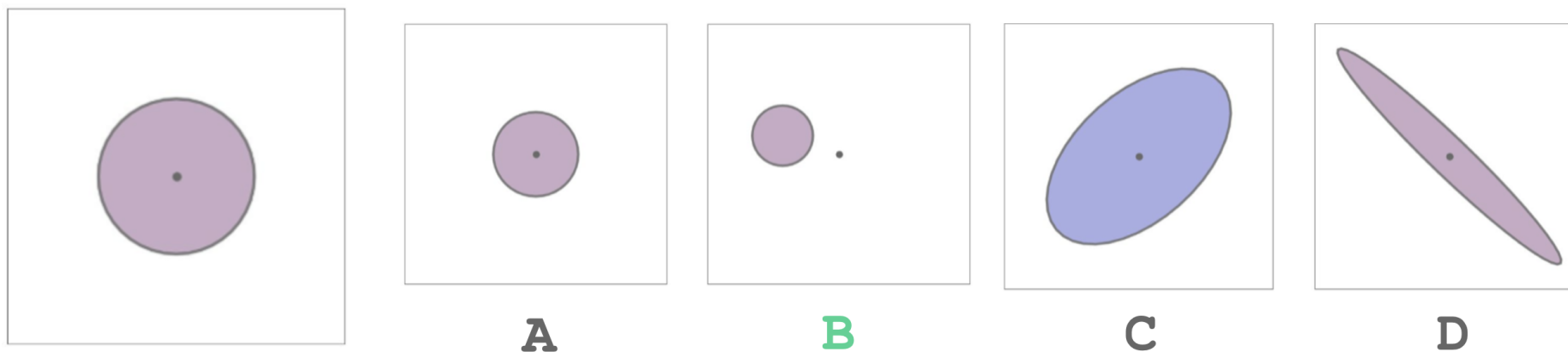
Integrated
Spatial Reasoning

Tangram Puzzle



Can the Tangram puzzle be completed with the available pieces, **yes** or no?

2D Transformation



Make it smaller then move it left and slightly upwards.

Foundational
Geometric
Transformations

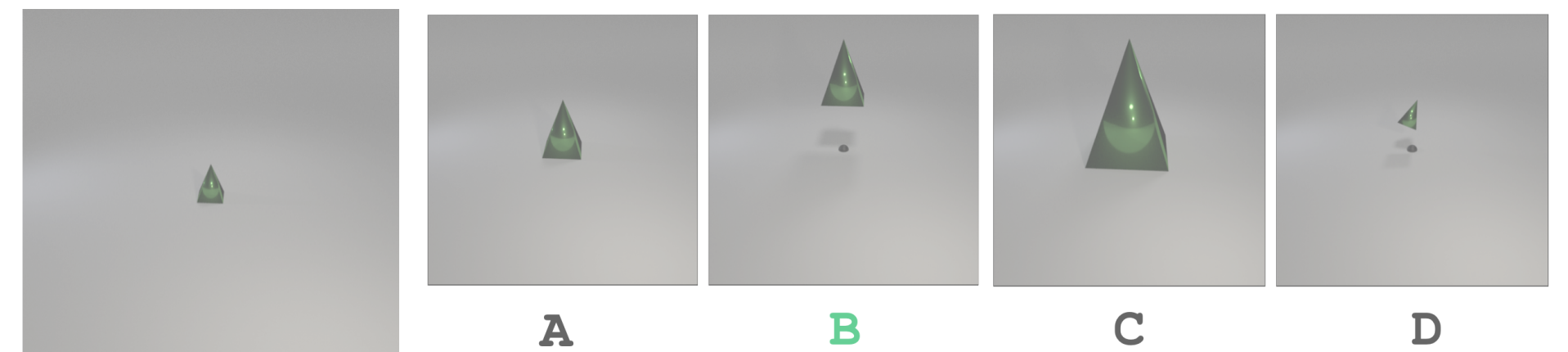
Cube Net Folding

Check out a net with 6 faces below:



Can the net be folded to form a cube, **yes** or no?

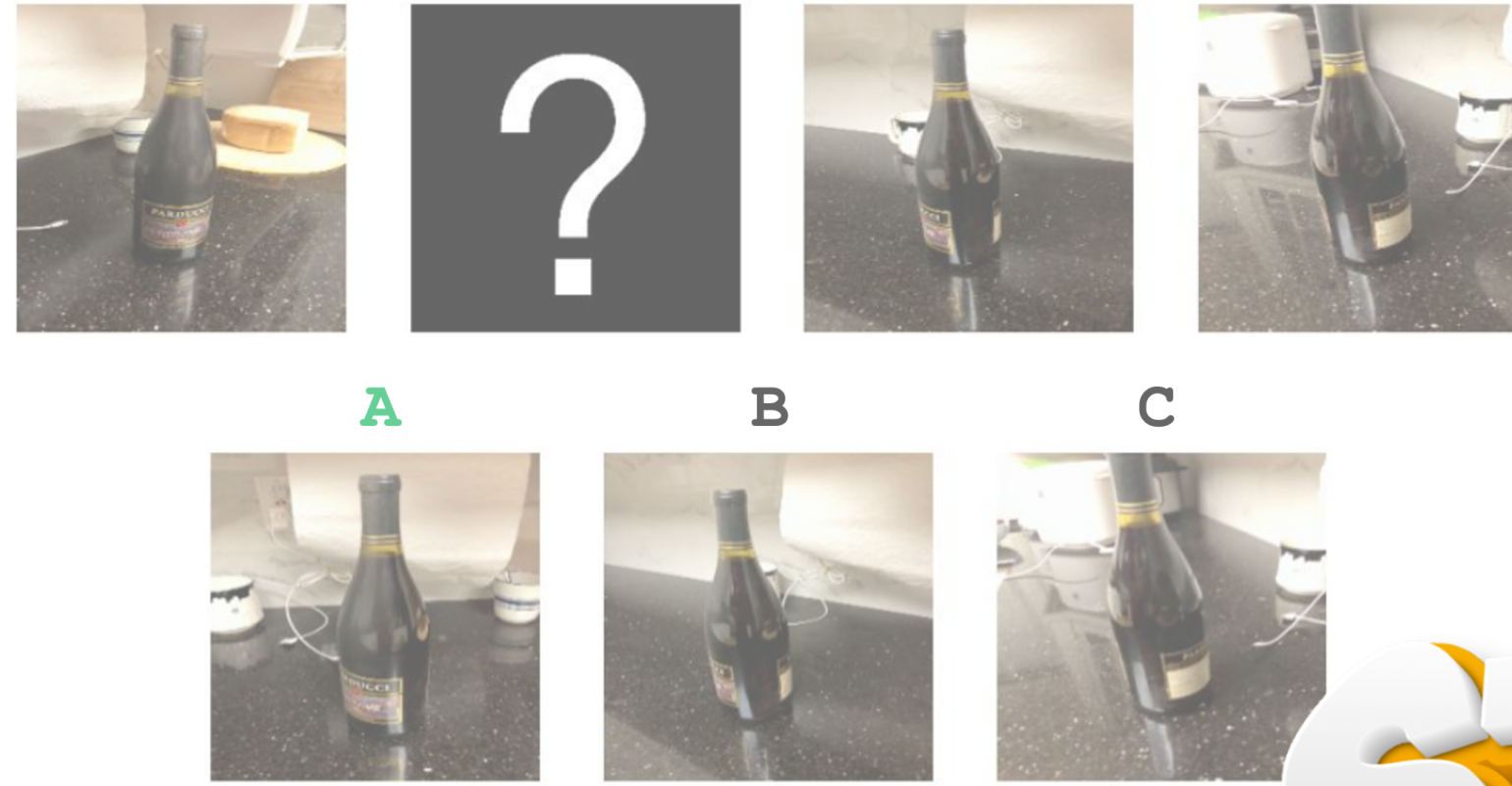
3D Transformation



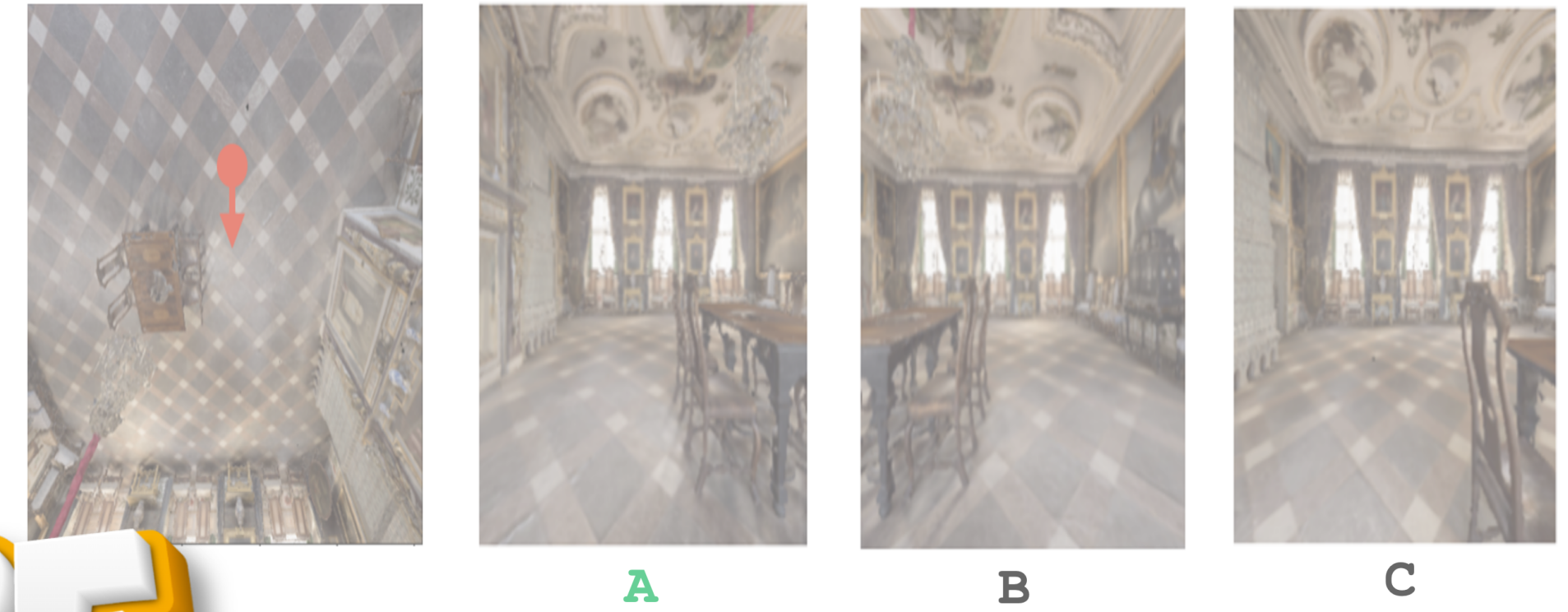
Move it up and then make it bigger.

Real-world
Spatial Reasoning

Temporal Frame Inference



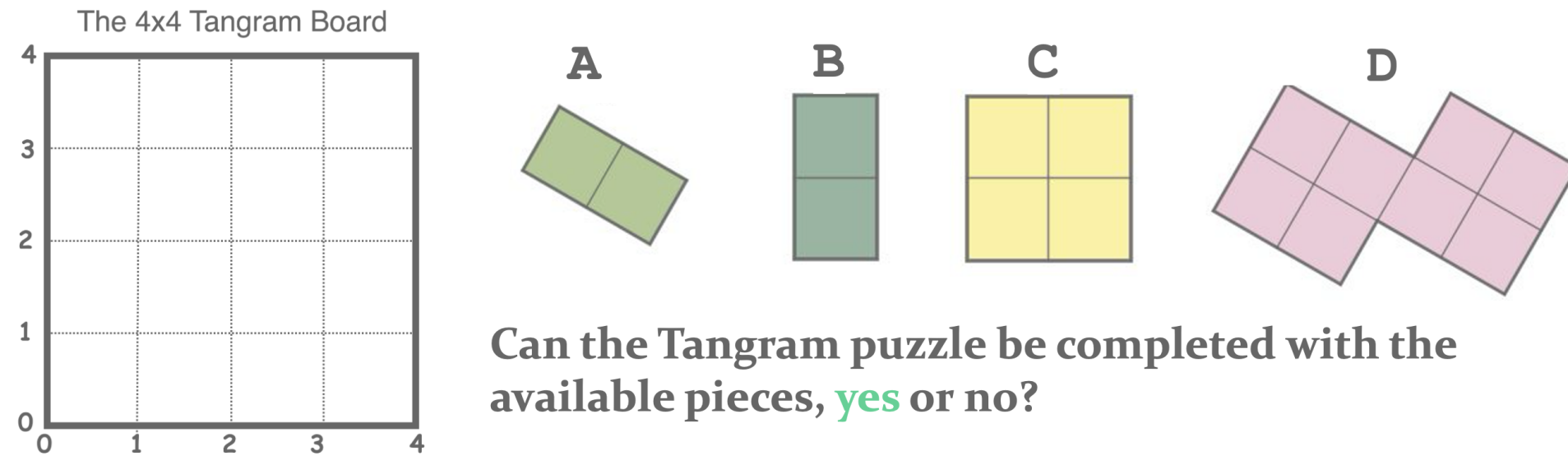
Perspective Reasoning



STARE

Integrated
Spatial Reasoning

Tangram Puzzle



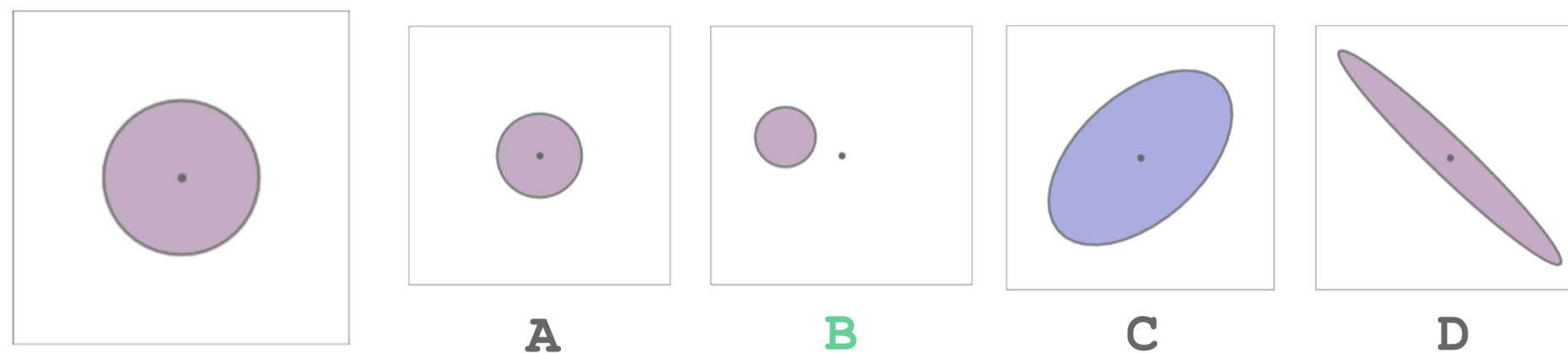
Cube Net Folding

Check out a net with 6 faces below:



Can the net be folded to form a cube, **yes** or no?

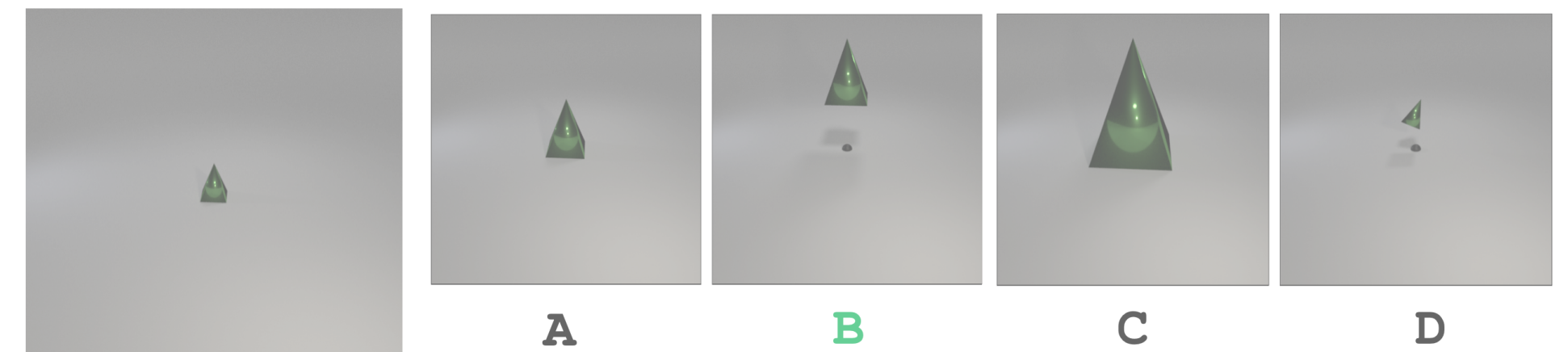
2D Transformation



Make it smaller then move it left and slightly upwards.

Foundational
Geometric
Transformations

3D Transformation



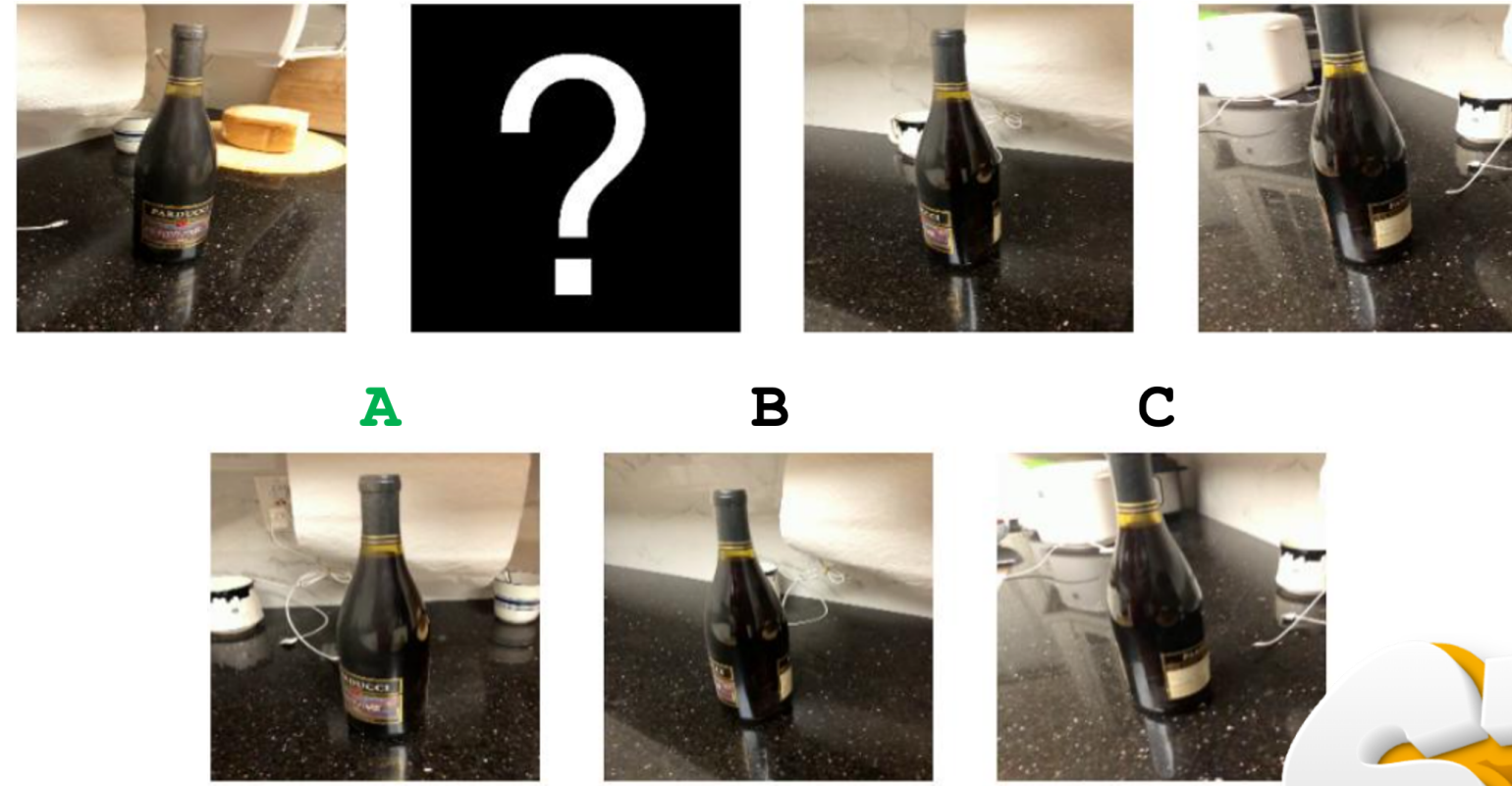
Move it up and then make it bigger.

"Implicit" Simulation

Temporal Frame Inference

Perspective Reasoning

Real-world
Spatial Reasoning

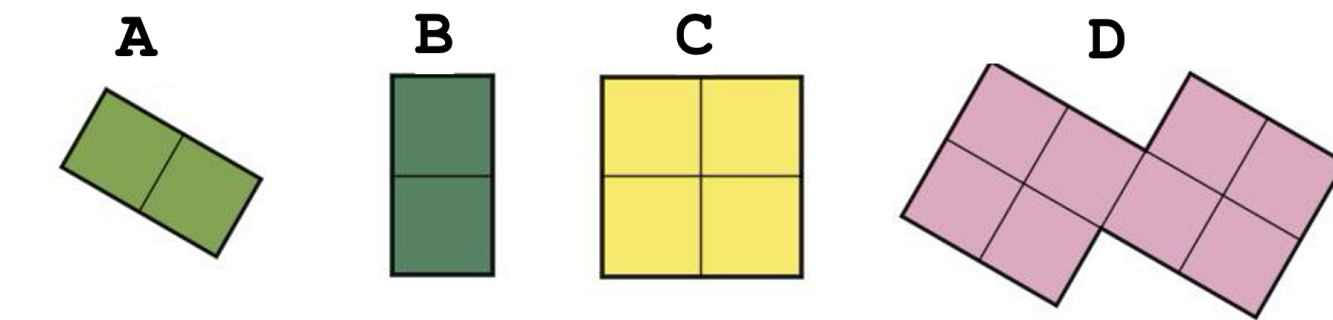
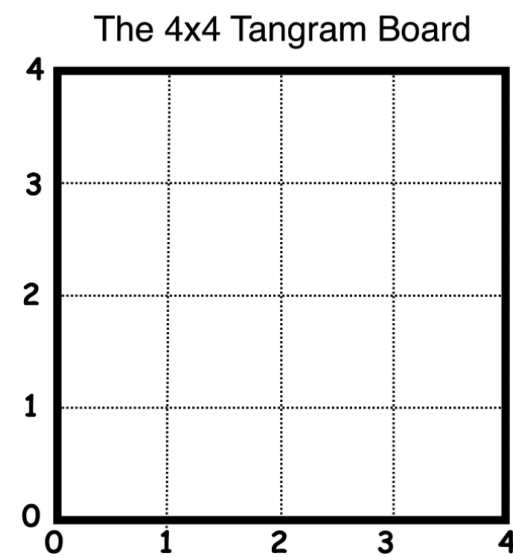


STARE

Tangram Puzzle

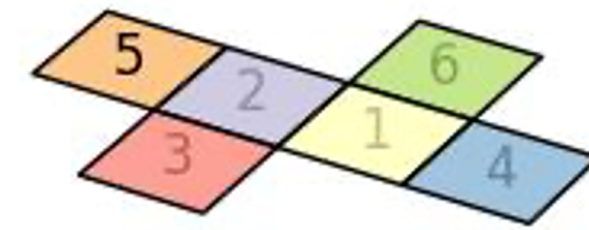
Cube Net Folding

Integrated
Spatial Reasoning



Can the Tangram puzzle be completed with the available pieces, **yes** or no?

Check out a net with 6 faces below:

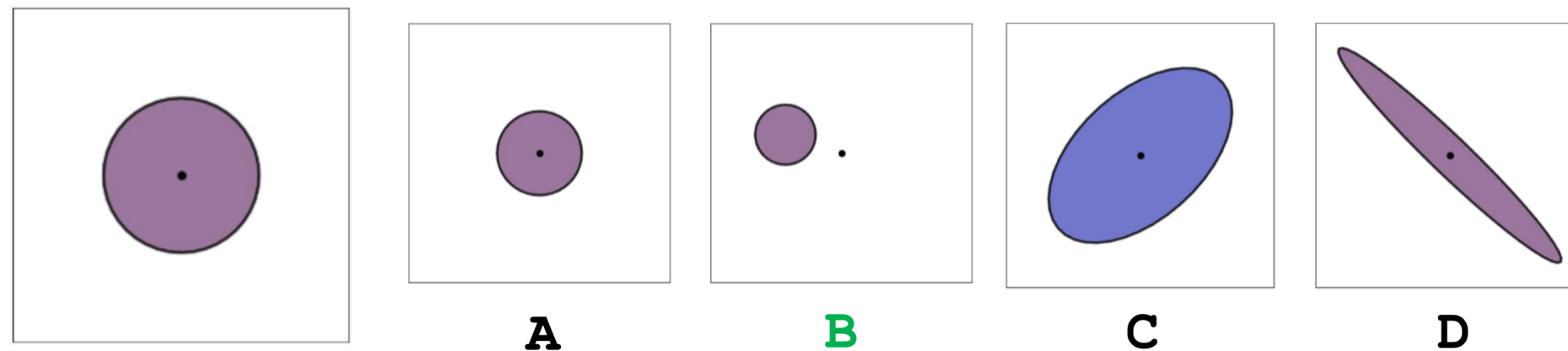


Can the net be folded to form a cube, **yes** or no?

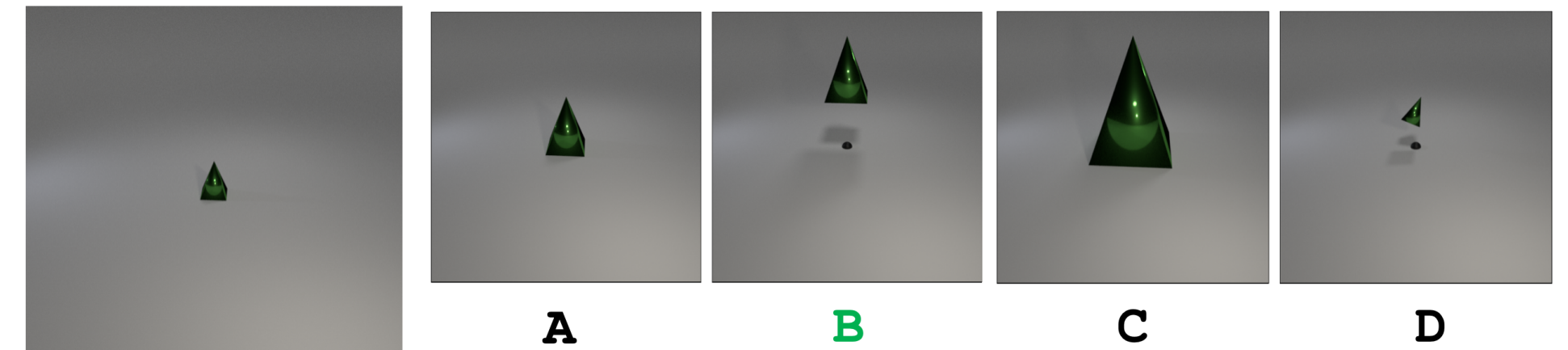
2D Transformation

3D Transformation

Foundational
Geometric
Transformations



Make it smaller then move it left and slig



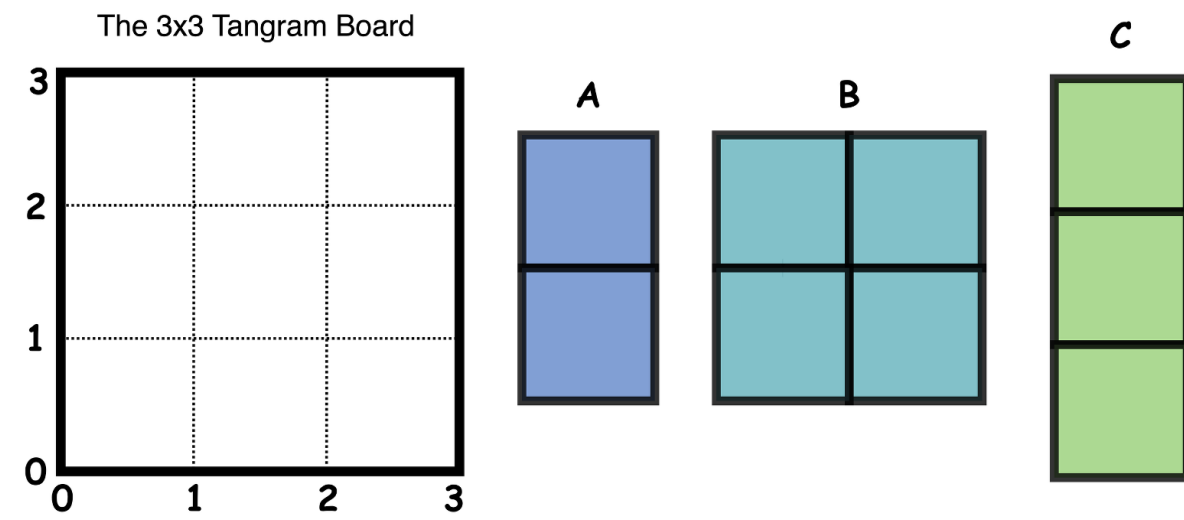
Move it up and then make it bigger.

"Explicit" Simulation

How to Evaluate Models on Visual Simulations? – “Explicit” Simulation

Question

Check out an Tangram puzzle below. The left panel is an empty Tangram puzzle, while the right panel shows available pieces to complete the puzzle.



Steps

Below are the steps to complete the Tangram puzzle:

Step 1: Rotate piece B by about 90 degrees clockwise, and place piece B with its upper-left corner at $(x, y) = (1, 3)$.

Step 2: Place piece A with its upper-left corner at $(x, y) = (0, 3)$.

Step 3: Rotate piece C by about 90 degrees clockwise, and place piece C with its upper-left corner at $(x, y) = (0, 1)$.

Based on the above steps, can the Tangram puzzle be completed with the available pieces, yes or no?

Without Visual Simulations

1

Question

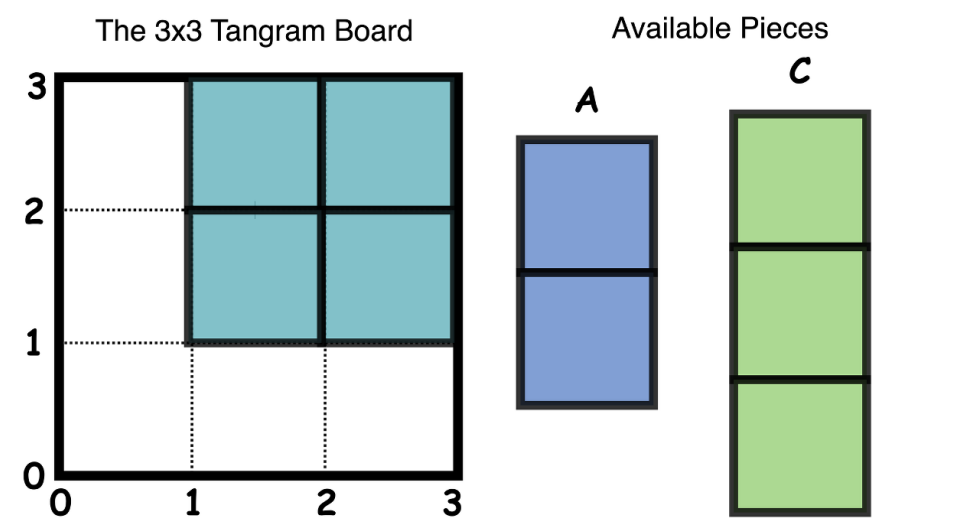
2

Question

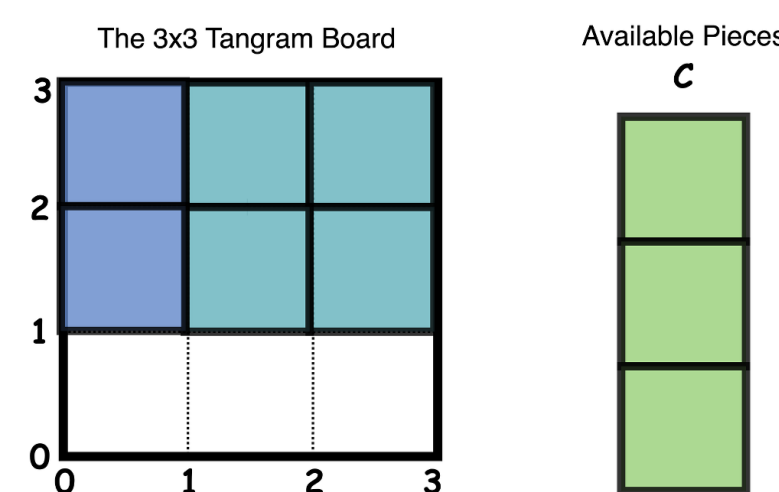
Steps

Intermediate Visual Simulations

1



2



With Visual Simulations

3

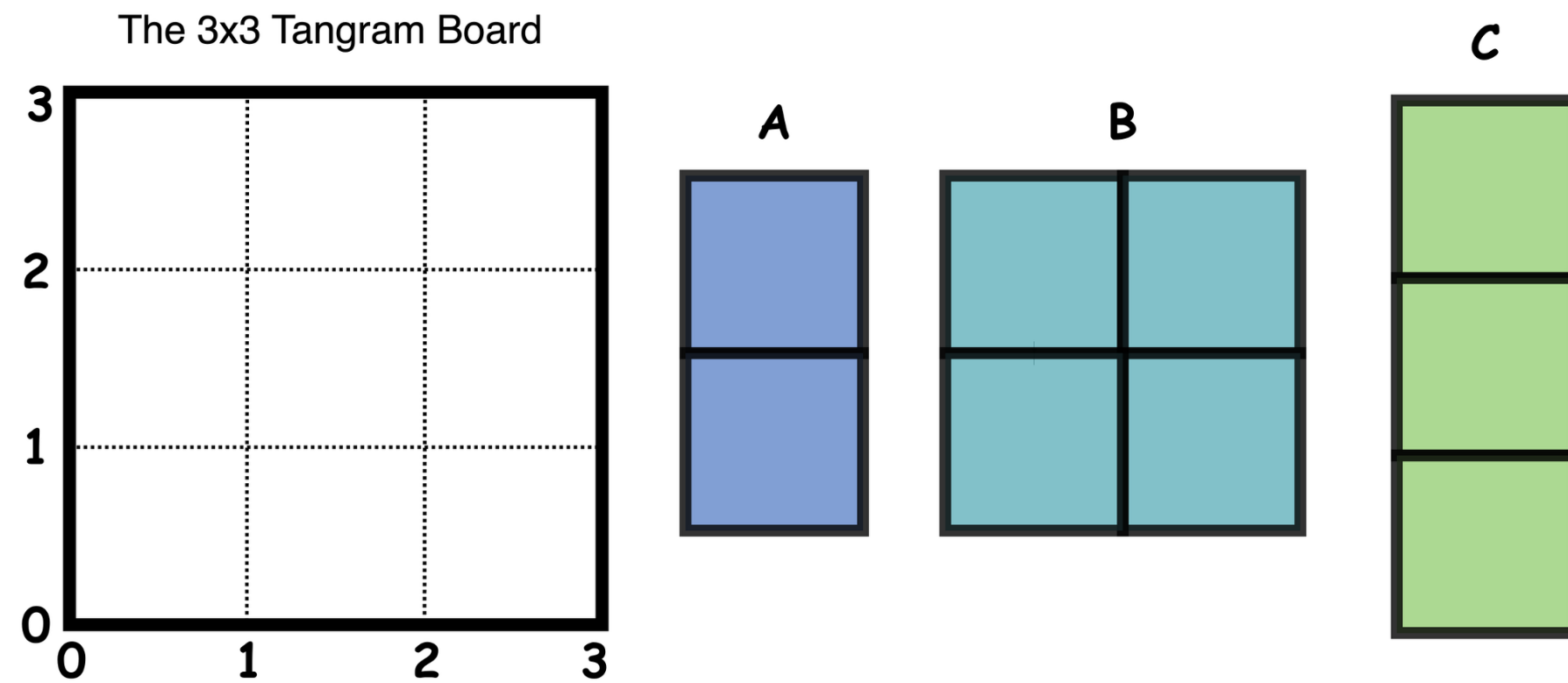
Question

Steps

Intermediate
Visual
Simulations

How to Evaluate Models on Visual Simulations? – “Explicit” Simulation

Check out an Tangram puzzle below. The left panel is an empty Tangram puzzle, while the right panel shows available pieces to complete the puzzle.



Without Visual Simulations

1

Question

Below are the steps to complete the Tangram puzzle:

Step 1: Rotate piece B by about 90 degrees clockwise, and place piece B with its upper-left corner at $(x, y) = (1, 3)$.

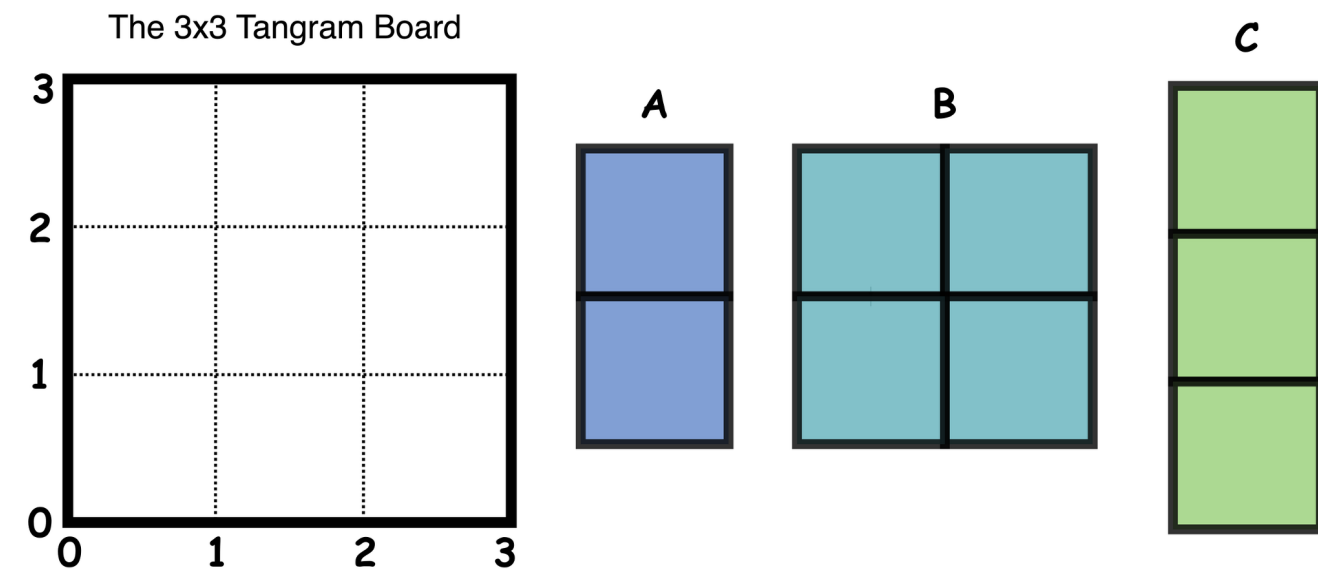
Step 2: Place piece A with its upper-left corner at $(x, y) = (0, 3)$.

Step 3: Rotate piece C by about 90 degrees clockwise, and place piece C with its upper-left corner at $(x, y) = (0, 1)$.

Based on the above steps, can the Tangram puzzle be completed with the available pieces, yes or no?

How to Evaluate Models on Visual Simulations? – “Explicit” Simulation

Check out an Tangram puzzle below. The left panel is an empty Tangram puzzle, while the right panel shows available pieces to complete the puzzle.



Without Visual Simulations

1

Question

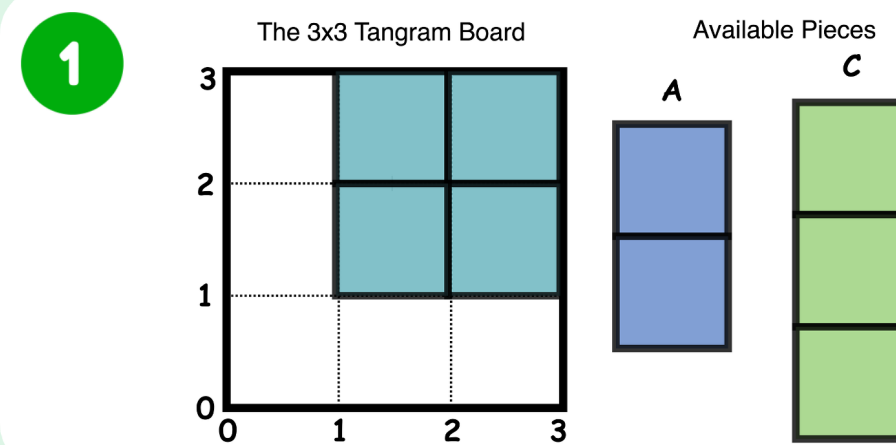
2

Question

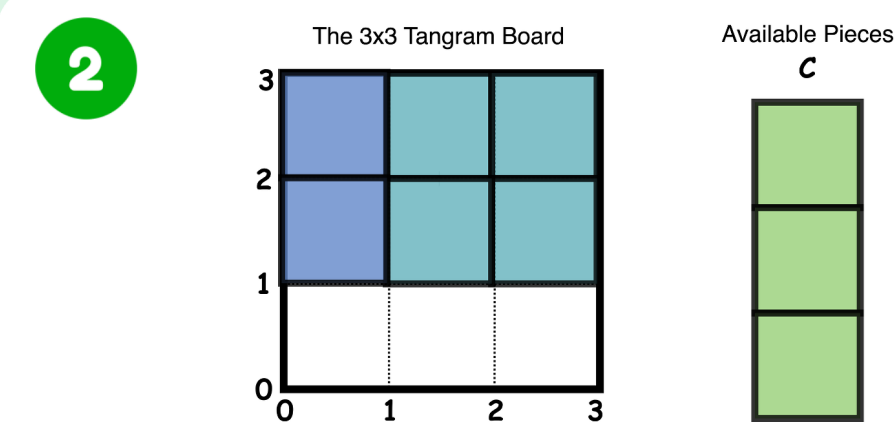
Steps

Below are the steps to complete the Tangram puzzle:

Step 1: Rotate piece B by about 90 degrees clockwise, and place piece B with its upper-left corner at $(x, y) = (1, 3)$.



Step 2: Place piece A with its upper-left corner at $(x, y) = (0, 3)$.



Step 3: Rotate piece C by about 90 degrees clockwise, and place piece C with its upper-left corner at $(x, y) = (0, 1)$.

Based on the above steps, can the Tangram puzzle be completed with the available pieces, yes or no?

With Visual Simulations

3

Question

Intermediate
Visual
Simulations

Steps

How to Evaluate Models on Visual Simulations? – “Implicit” Simulation

Question

You see 4 sequential frames of a video, but one is missing (marked with "?"). Choose which of the images in the second row correctly fills the missing frame. Remember, the camera only moves in one direction (left or right) in the video.



Choose the correct missing frame.



A

B

C

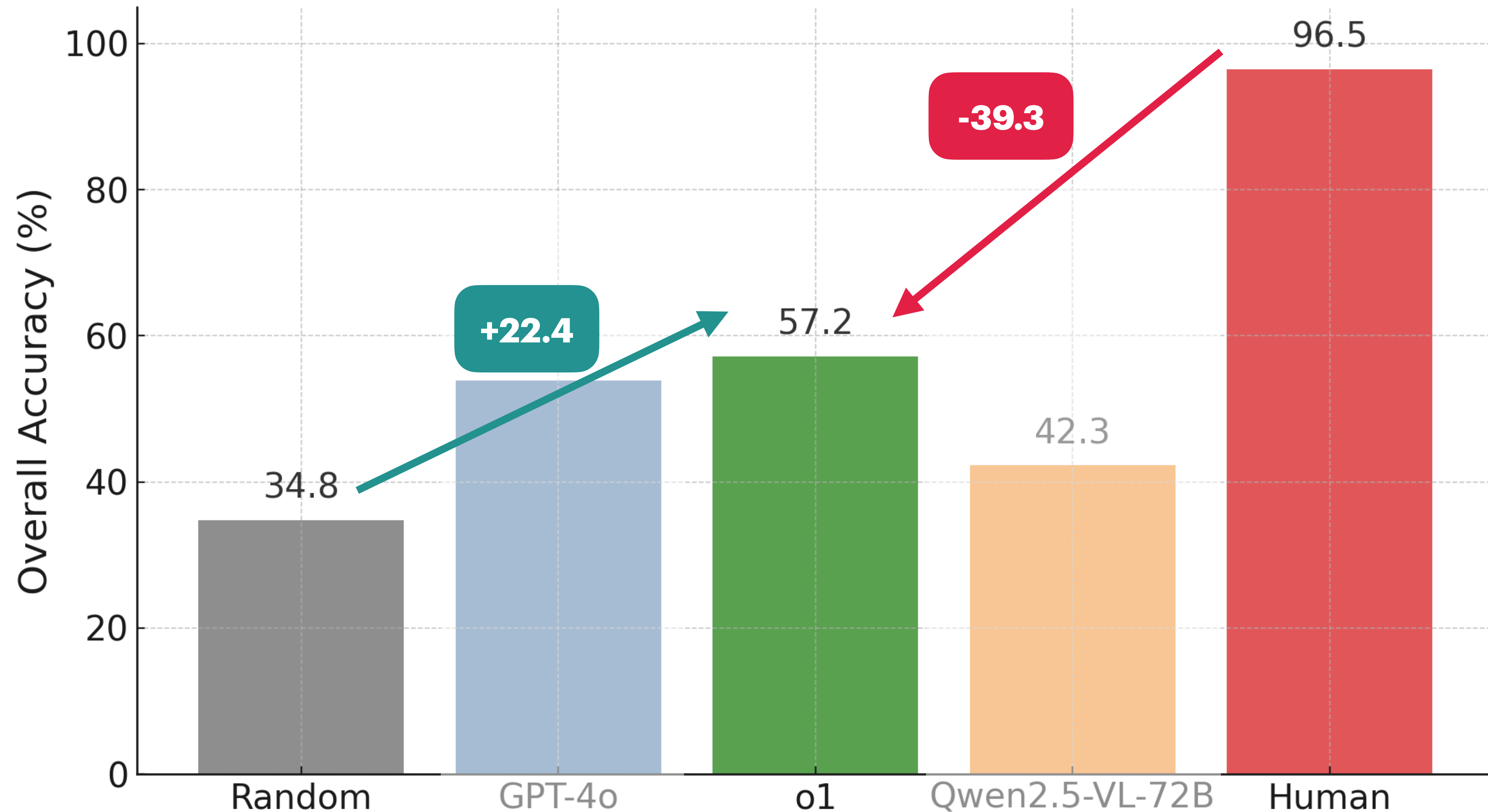
Without Visual Simulations

1

Question

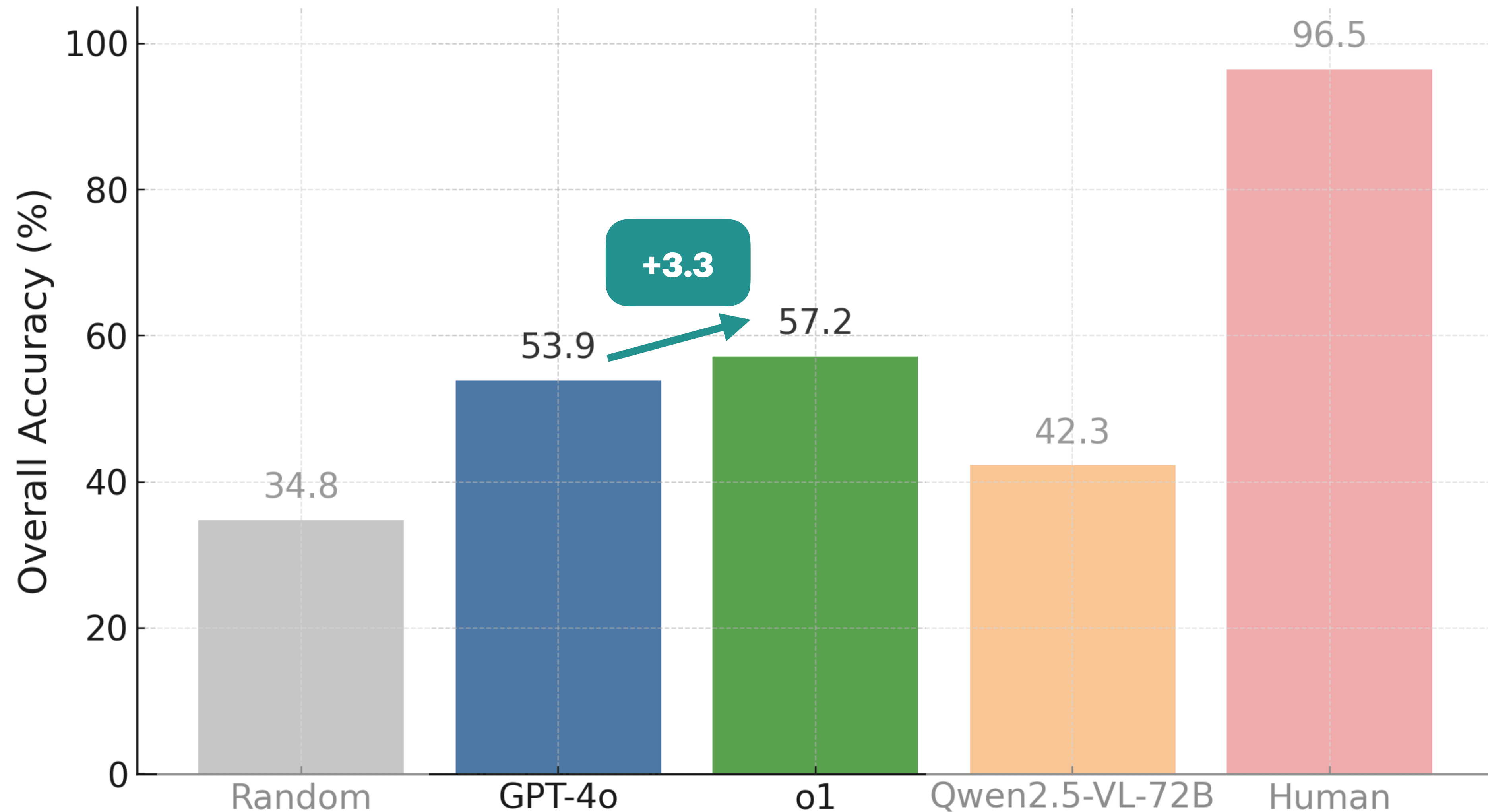
How Well do Models Perform on STARE?

Human-model gap remains large!



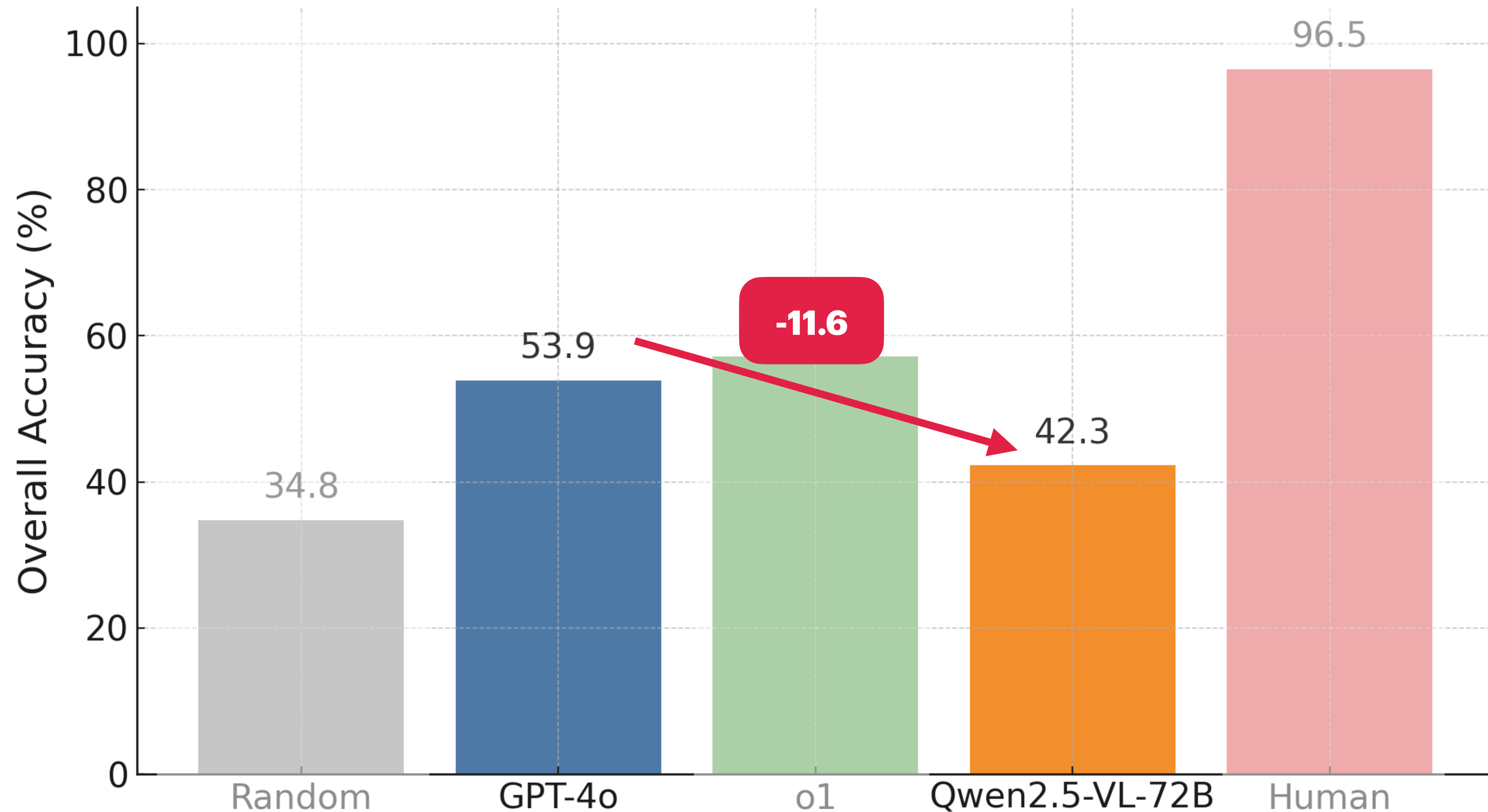
How Well do Models Perform on STARE?

Reasoning-focused o1 only improved GPT-4o by a few points.



How Well do Models Perform on STARE?

Open-source model still lag behind the proprietary models.

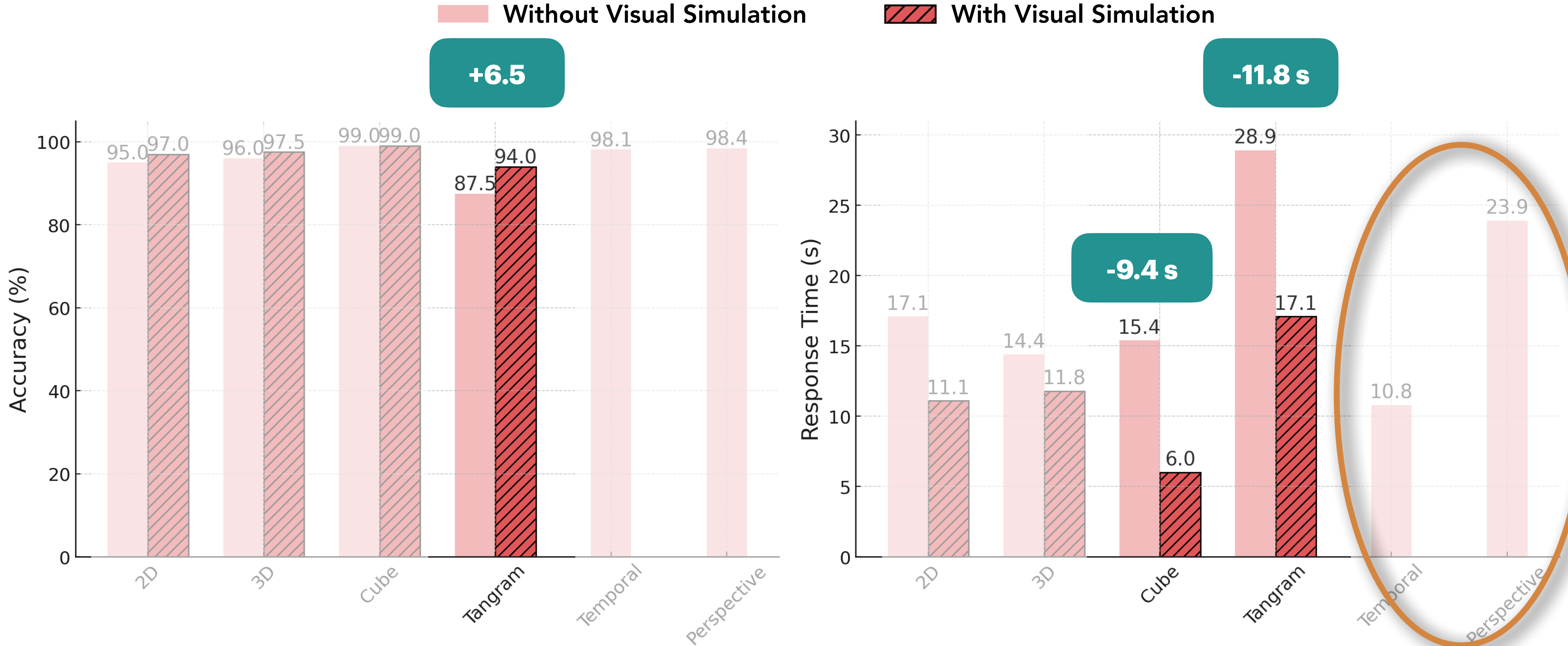


Are Visual Simulations Helpful? — for



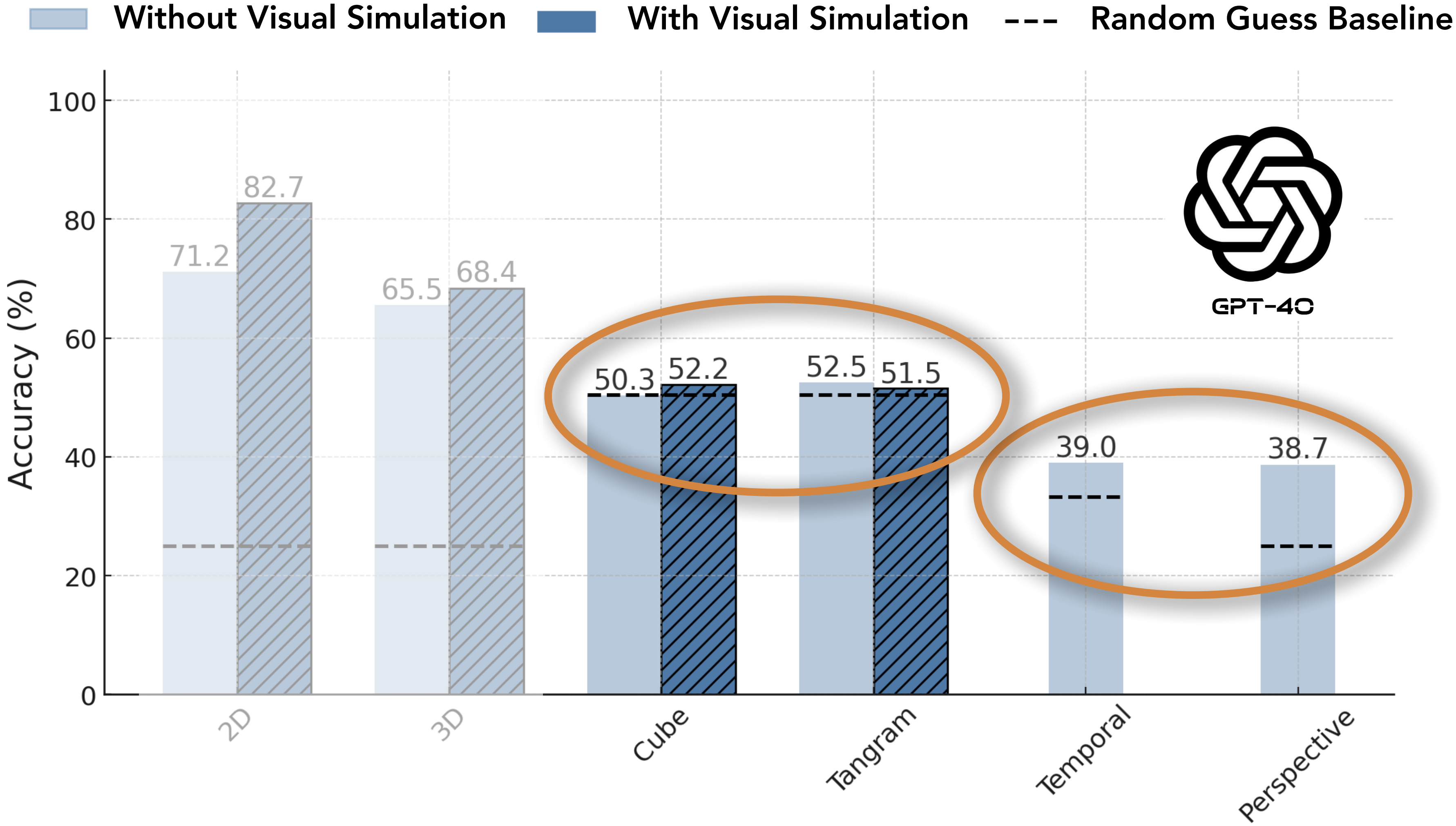
Human

Small gains in performance but significant decrease in response time.



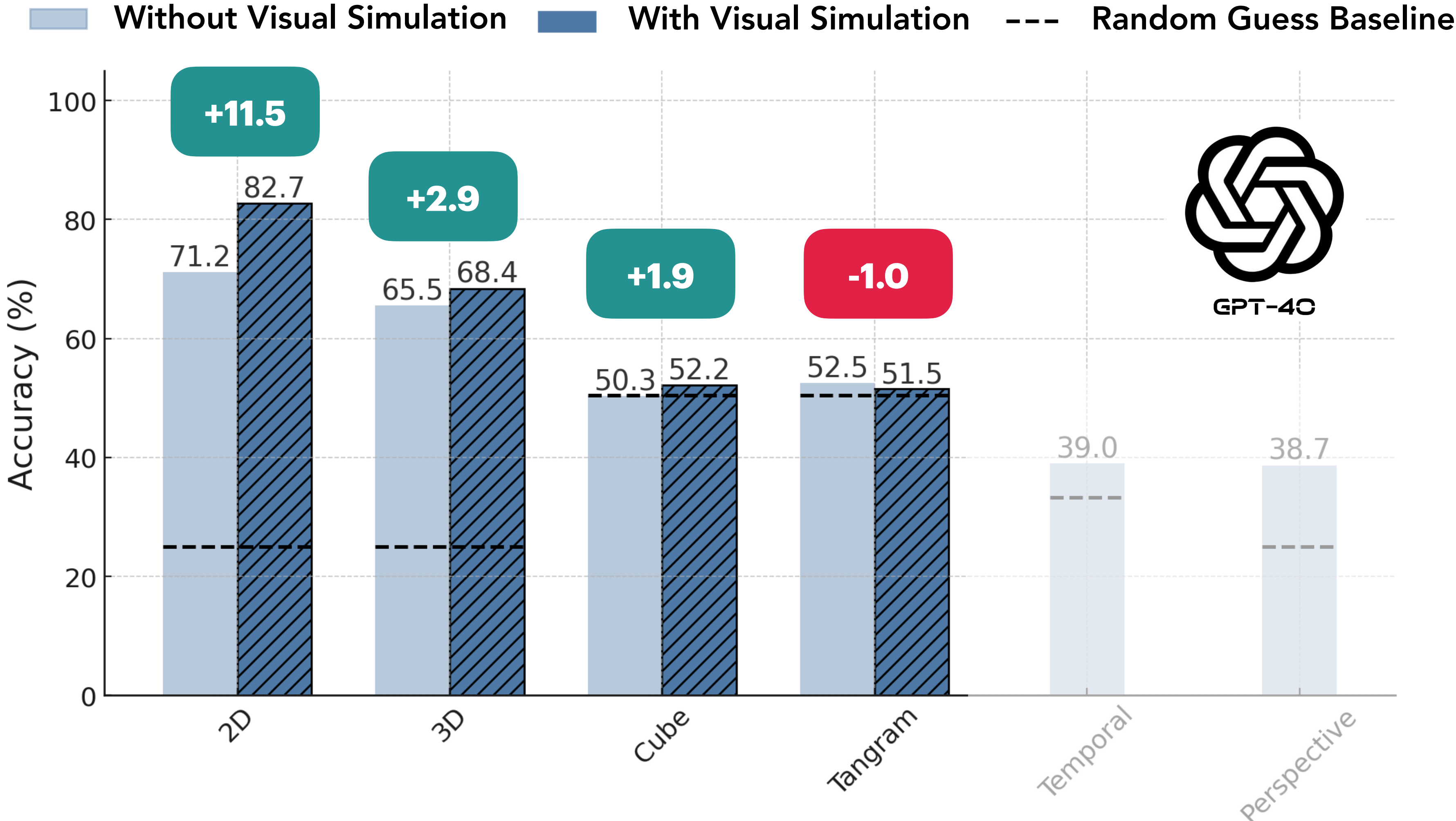
Are Visual Simulations Helpful? — for Models

GPT-4o performs close-to-random on harder tasks.



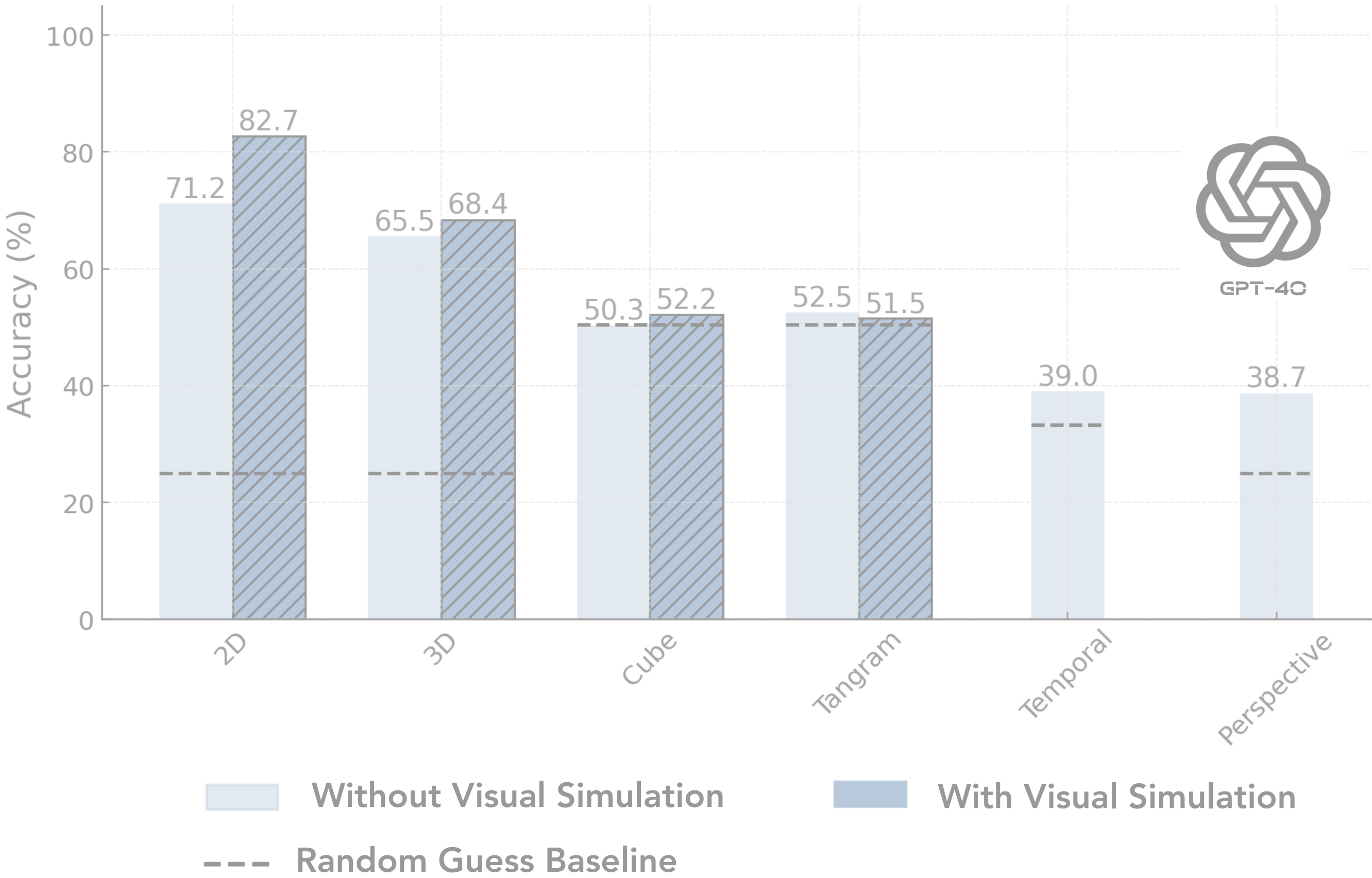
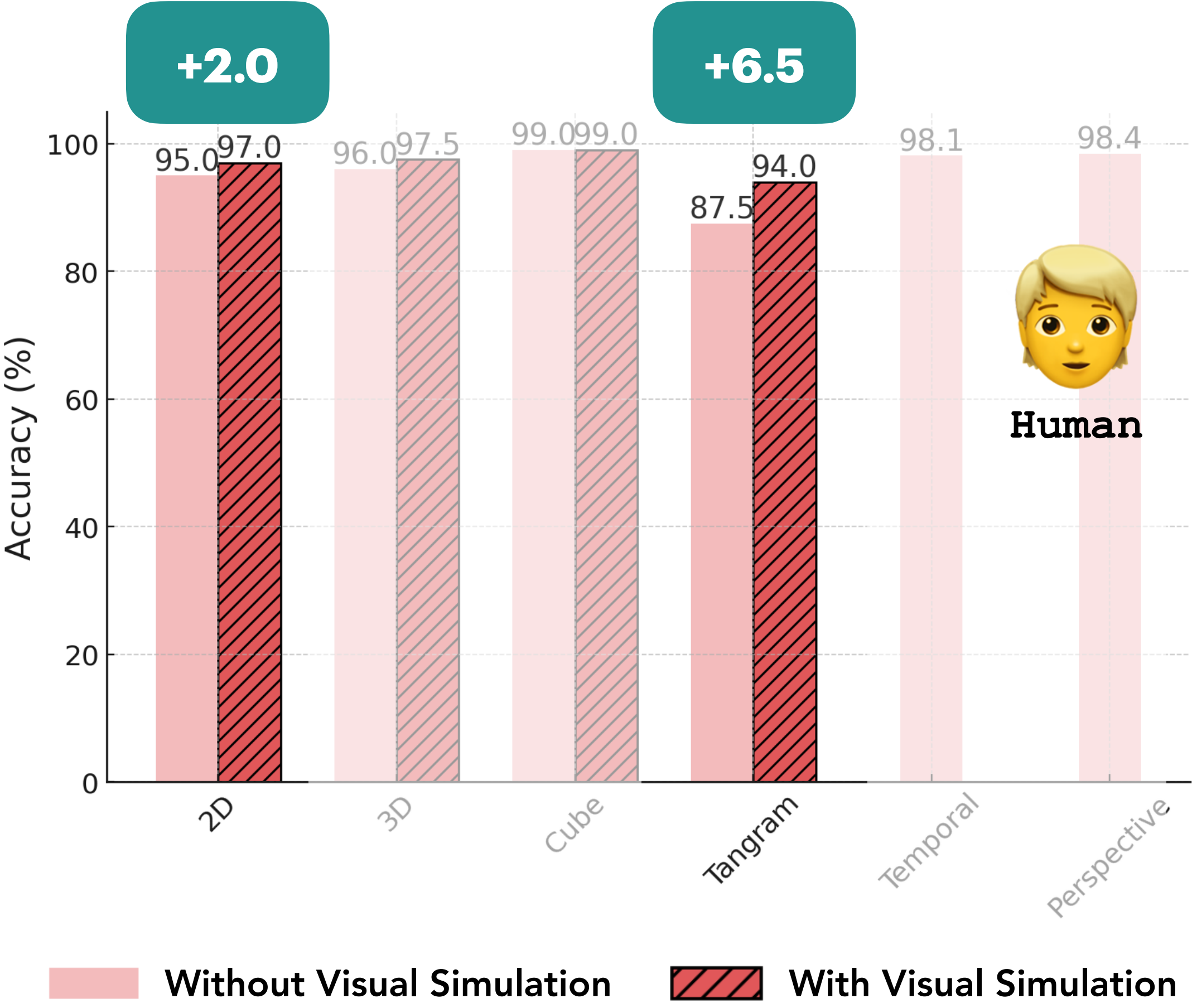
Are Visual Simulations Helpful? — for Models

Big gains for easier tasks (e.g., 2D Transformation), performance even worsen for harder tasks (e.g., Tangram Puzzles)



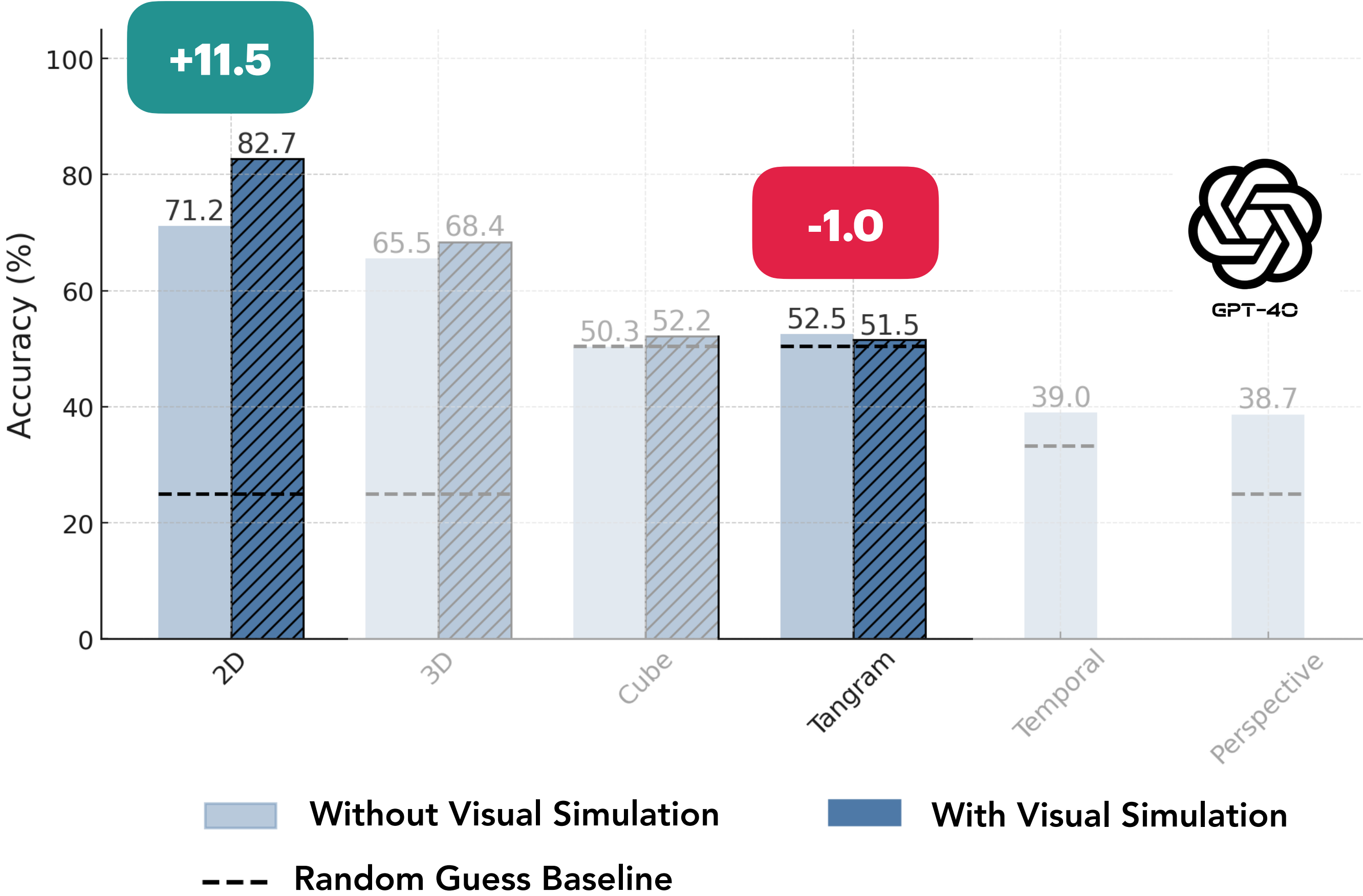
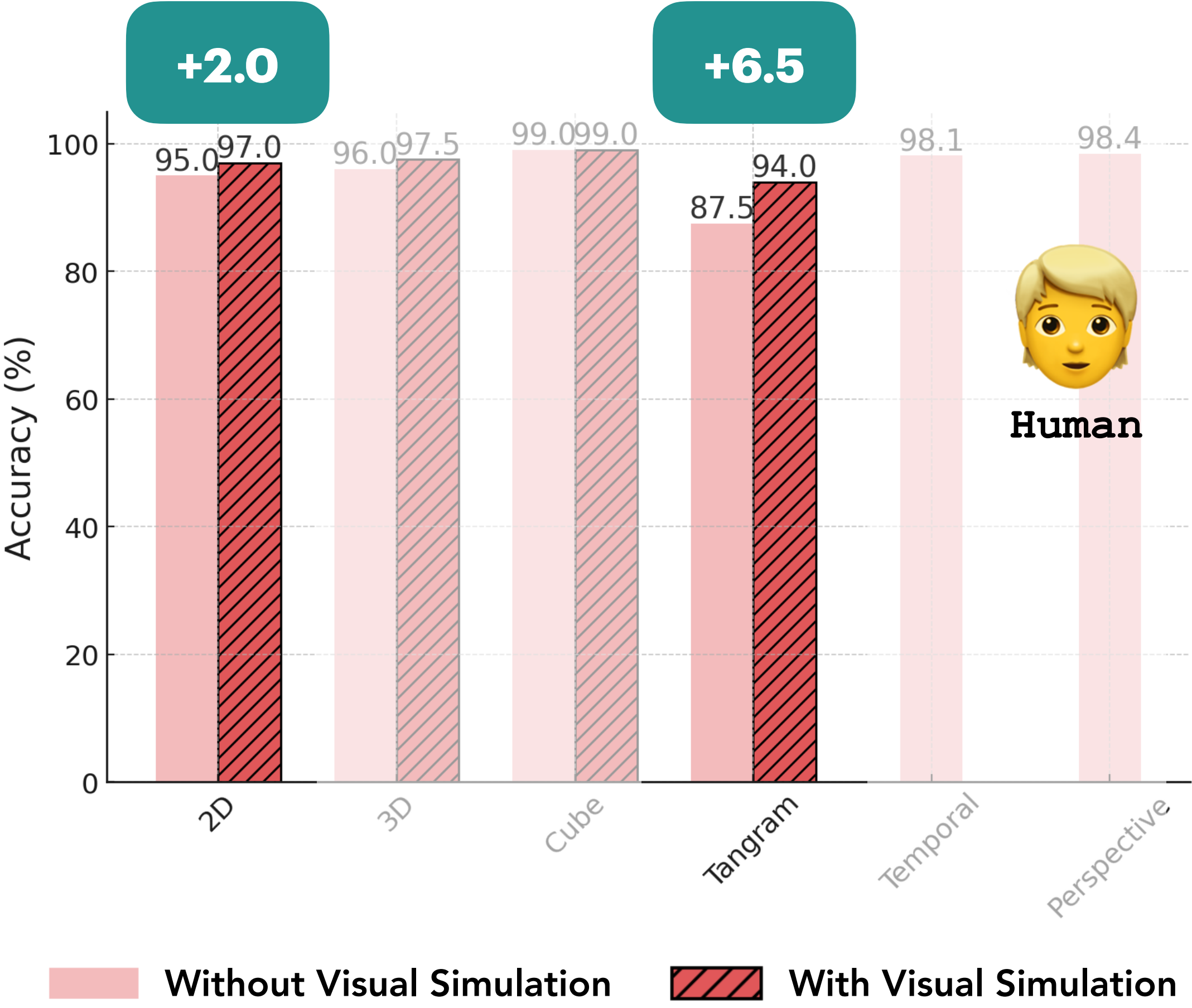
Are Visual Simulations Helpful? – Humans vs. Models

Humans benefit from visual simulation more on harder tasks than easier tasks



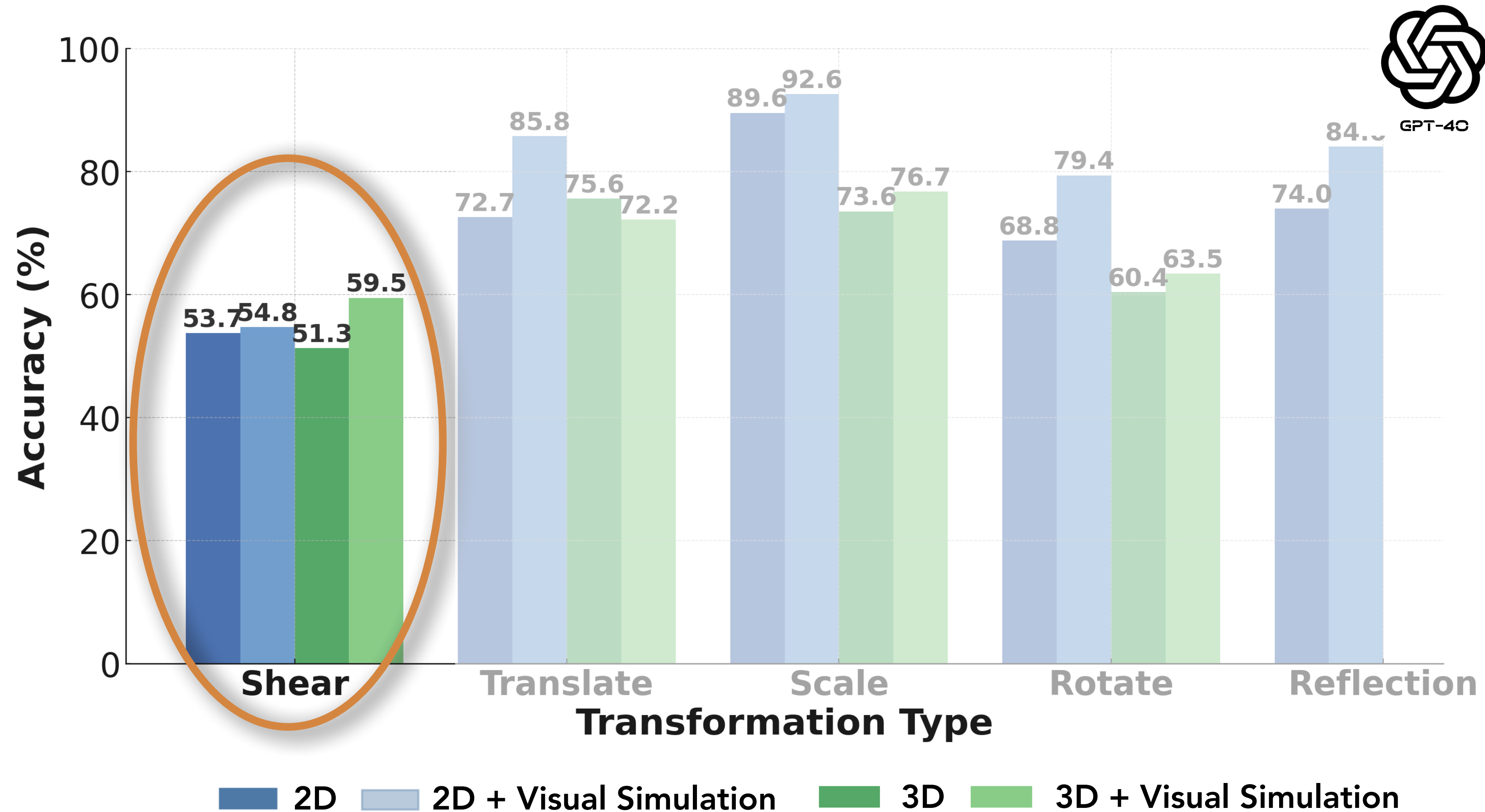
Are Visual Simulations Helpful? – Humans vs. Models

Humans benefit from visual simulation more on harder tasks than easier tasks, while models benefits more on easier tasks.



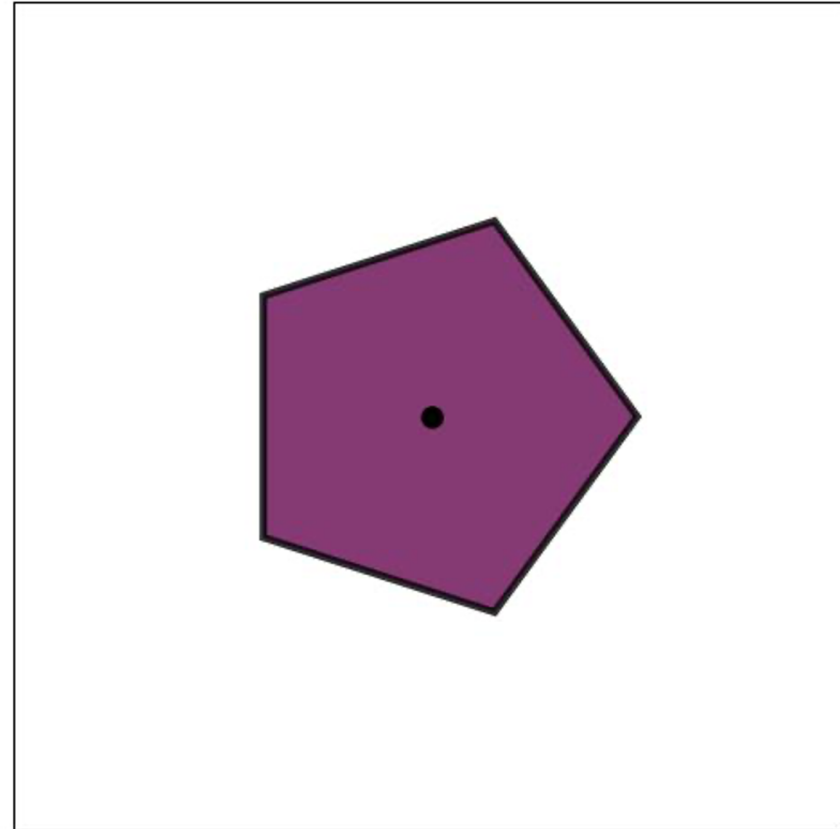
Do Models struggle with Basic Transformations?

Shearing is the most challenging transformation type for both 2D and 3D



How Much do Models Rely on “Easy” Perceptual Cues?

Easy



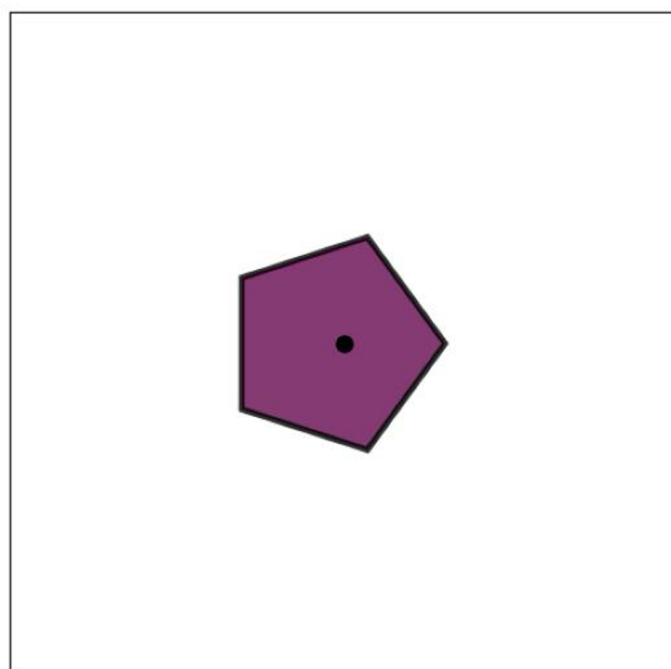
Imagine transforming a pentagon step by step.

Follow these changes:

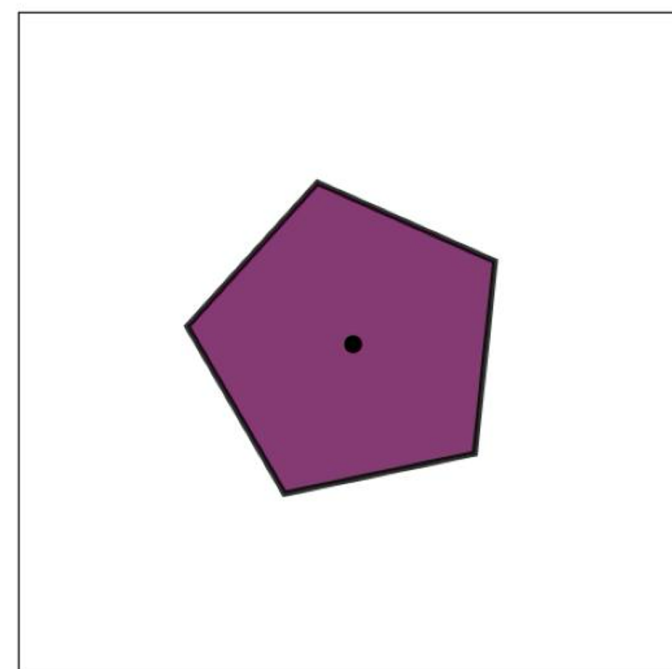
- Rotate the pentagon by 30 degrees counter-clockwise around its center.

After these transformations, which of the following shapes best represents the final result? For reference, the black dots in each panel of the figures indicate the origin.

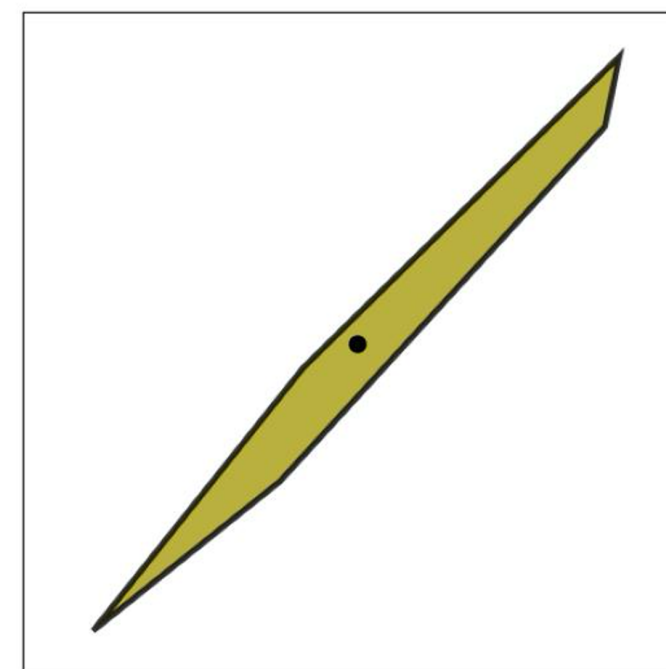
(A)



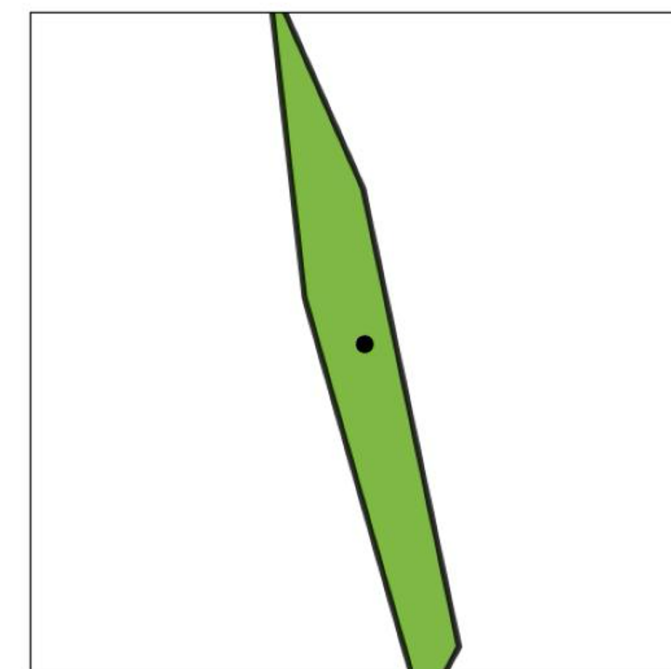
(B)



(C)

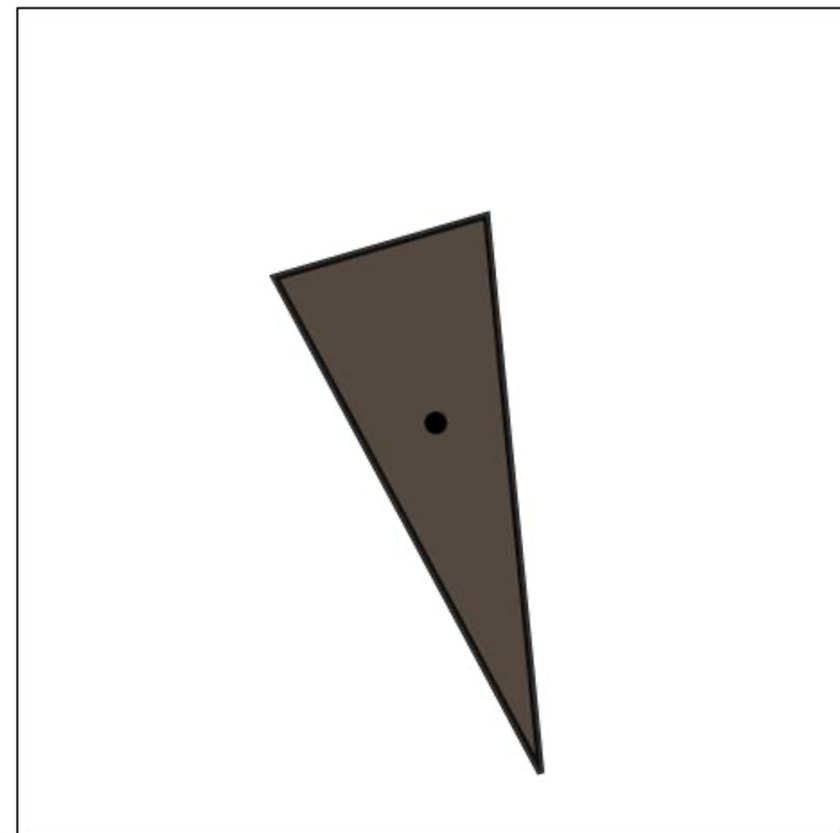


(D)



How Much do Models Rely on “Easy” Perceptual Cues?

Medium



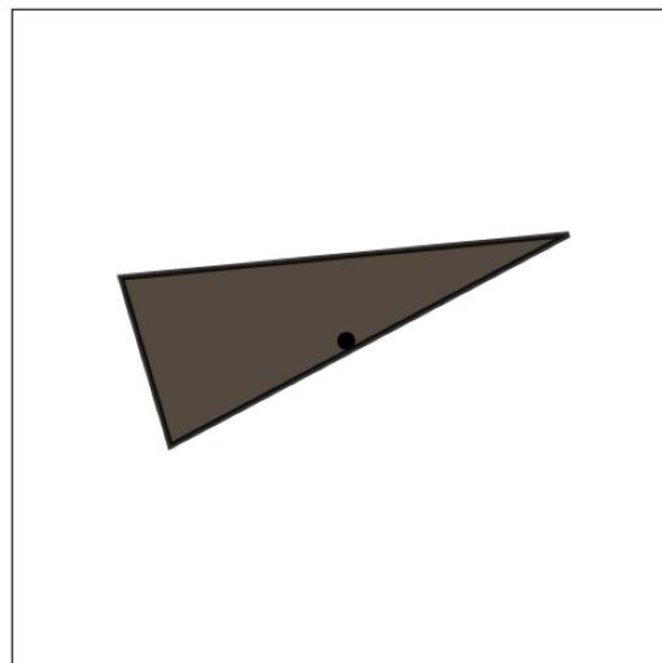
Imagine transforming a triangle step by step.

Follow these changes:

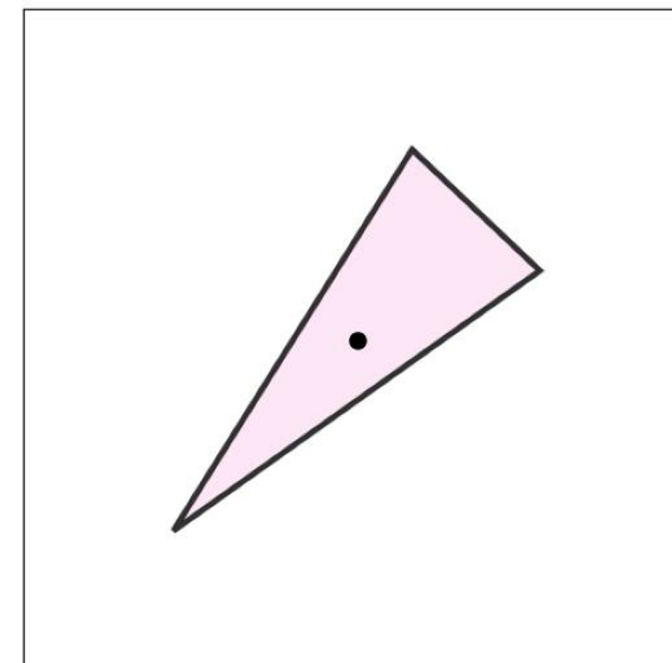
- Reduce the size of the triangle, making it smaller.

After these transformations, which of the following shapes best represents the final result? For reference, the black dots in each panel of the figures indicate the origin.

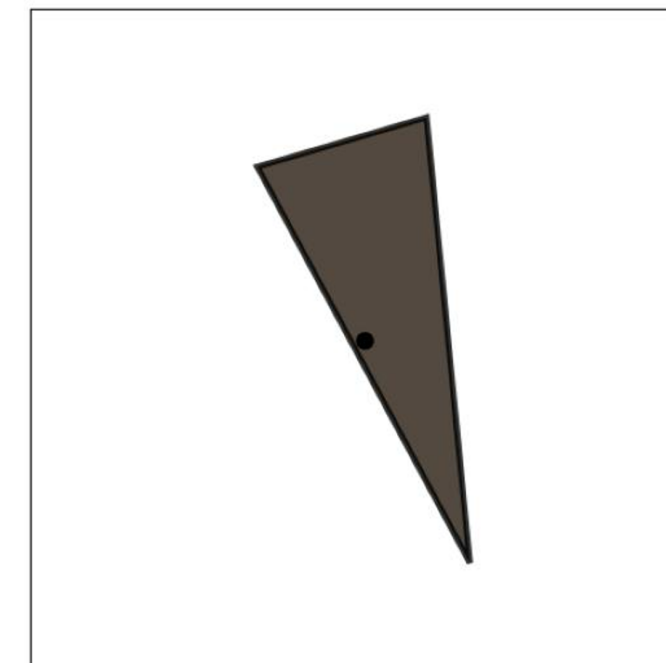
(A)



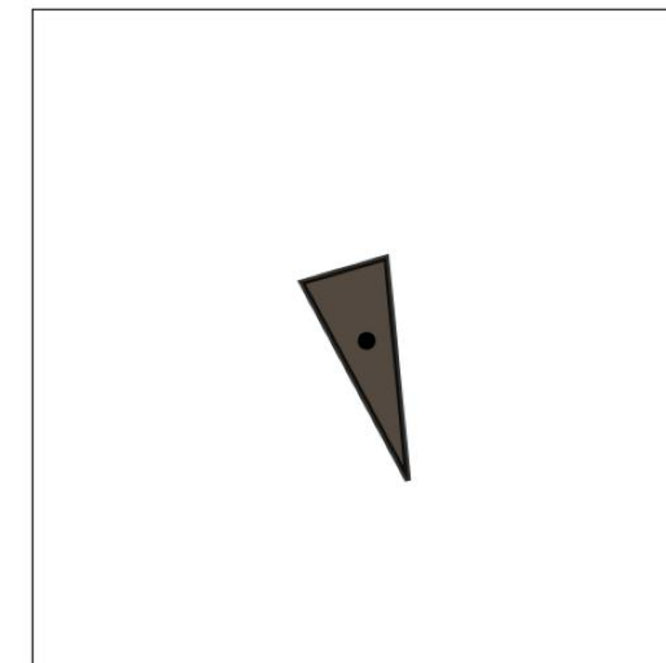
(B)



(C)

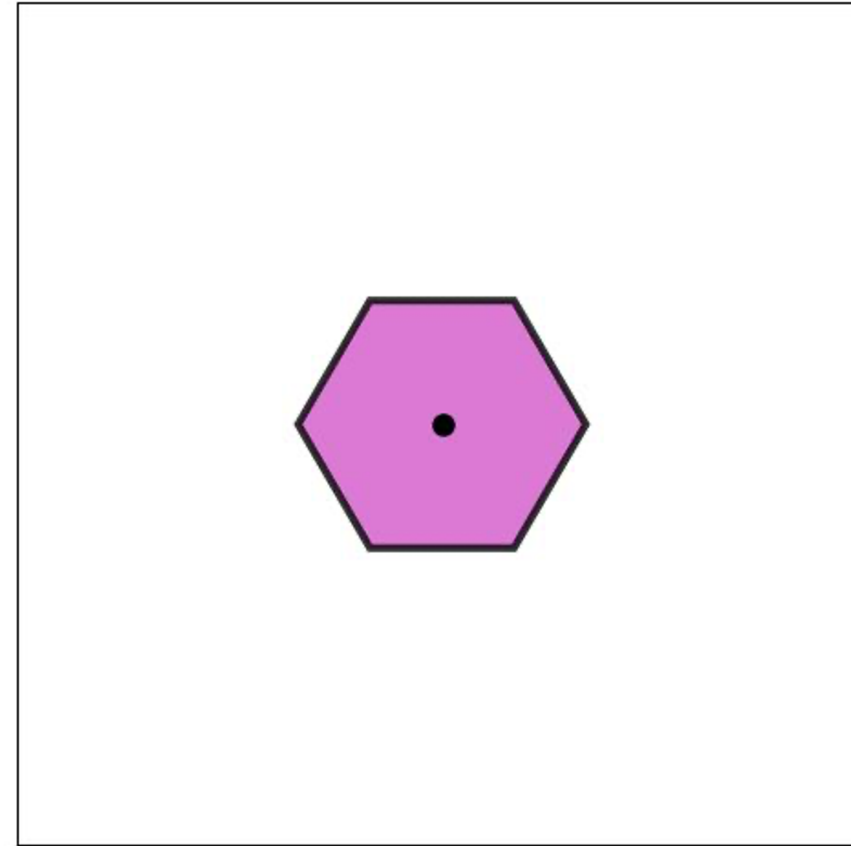


(D)



How Much do Models Rely on “Easy” Perceptual Cues?

Hard



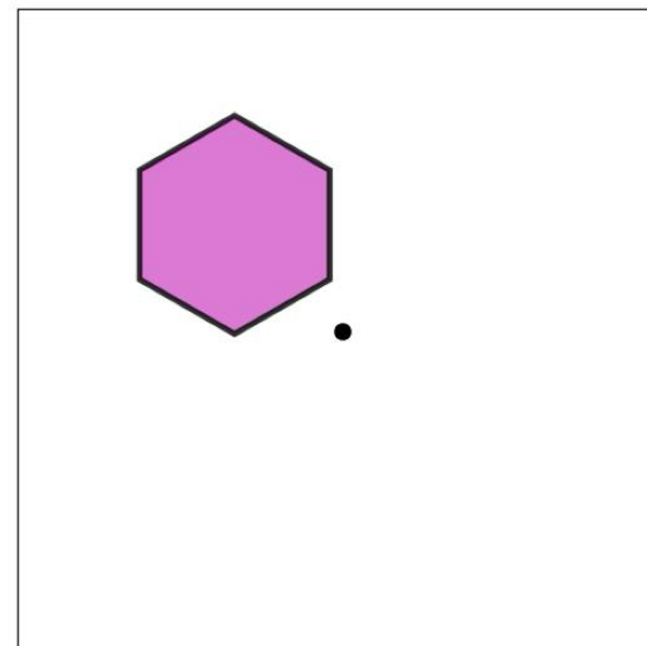
Imagine transforming a hexagon step by step.

Follow these changes:

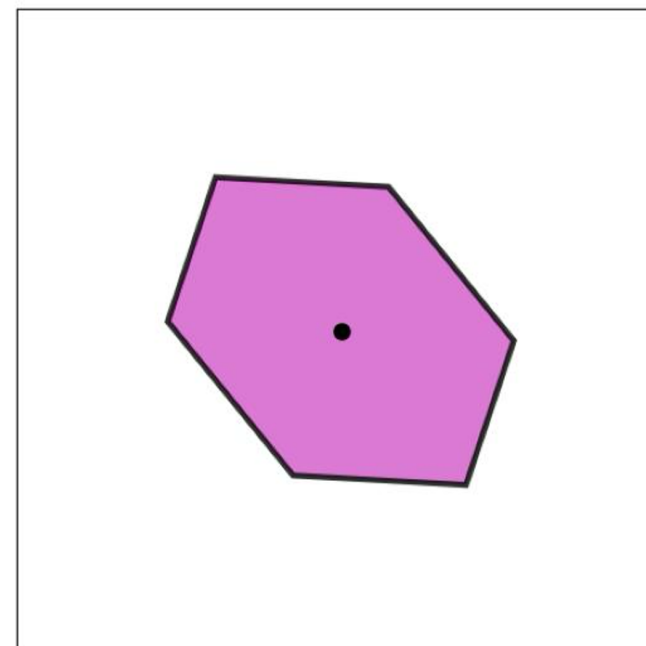
- Rotate the hexagon by 30 degrees clockwise around its center
- Shift the hexagon to the left by a significant amount and significantly upward.

After these transformations, which of the following shapes best represents the final result? For reference, the black dots in each panel of the figures indicate the origin.

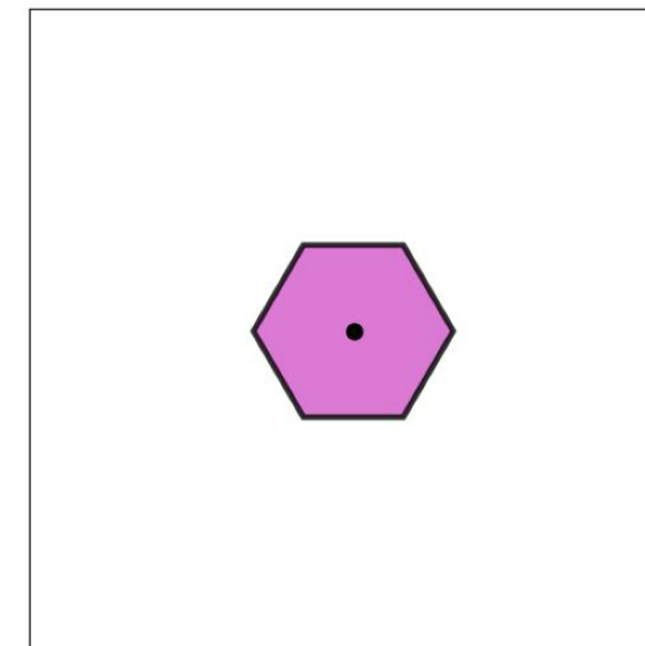
(A)



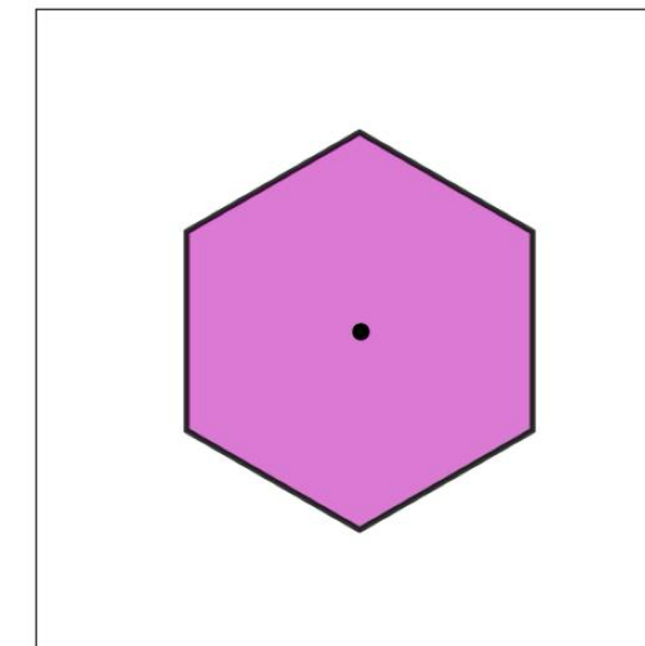
(B)



(C)

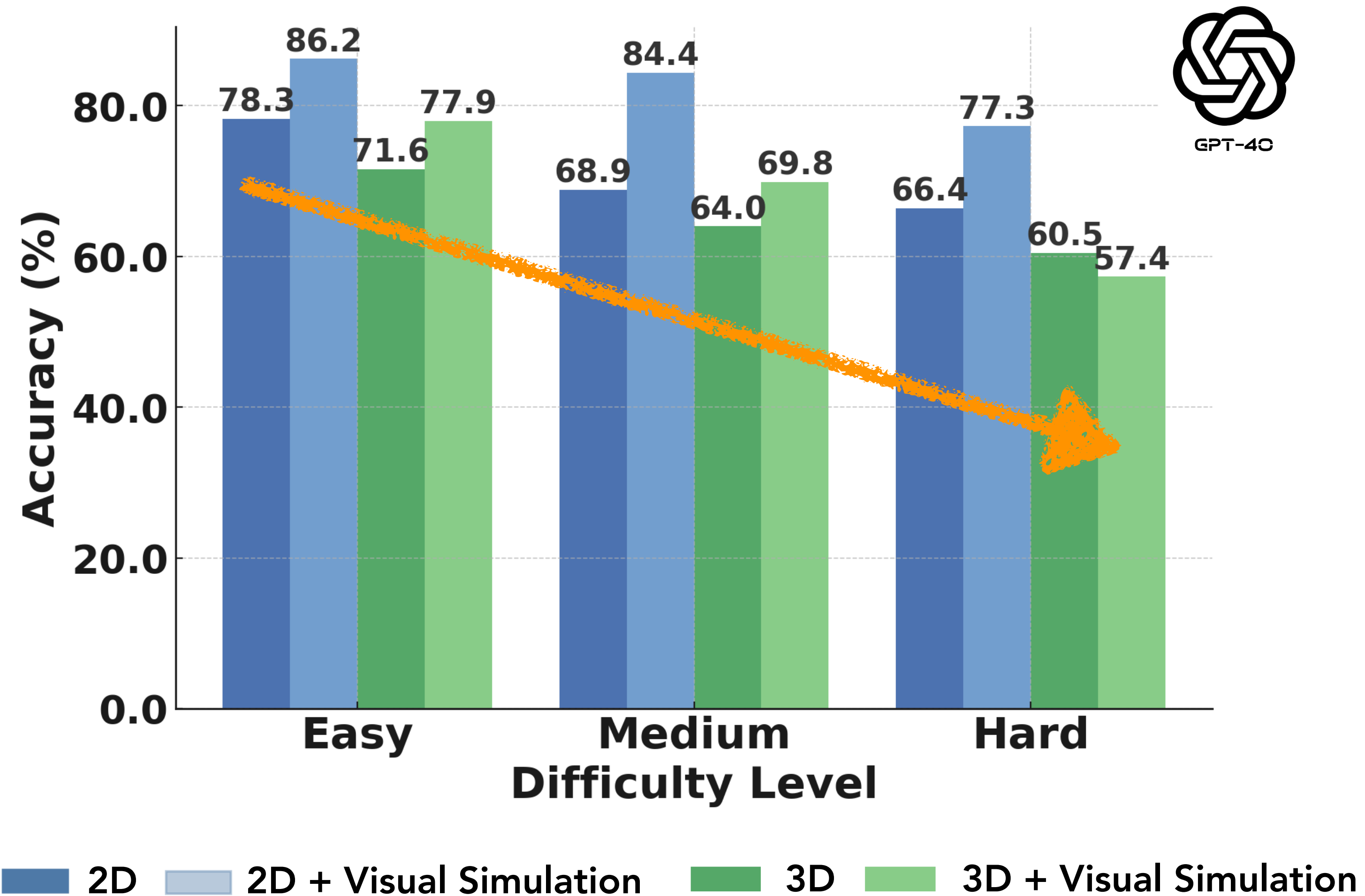


(D)



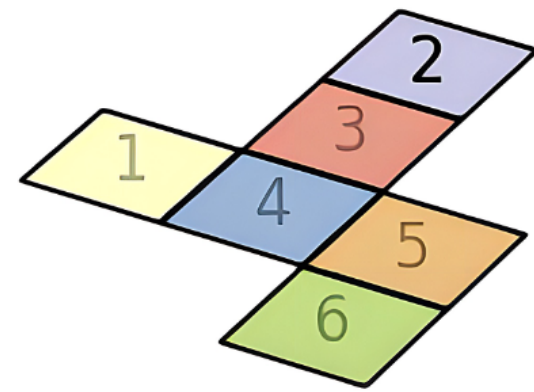
How Much do Models Rely on “Easy” Perceptual Cues?

Model performance drop as we reduce the number of easily distinguishable distractors.



Models are Not Perfect at Perception

Perception Error



Question: Check out a net with 6 square faces. Can the net be folded to form a cube, yes or no?

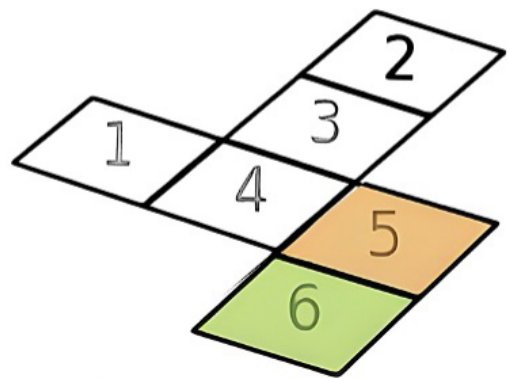
Claude:

Looking at the image, I see a net with 6 labeled squares arranged in a specific pattern:

- Square 1 (yellow) is on the left
 - Square 2 (purple) is at the top
 - Square 3 (pinkish) is in the middle right of Square 2
 - Square 4 (blue) is below Square 3 and to the right of Square 1
 - Square 5 (orange) is to the right of Square 4
 - **Square 6 (green) is below Square 4** ⚠
- So the answer is no, this specific net cannot be folded to form a cube.

Error Analysis:

Square 6 (green) is below Square 5.



2D Perception (Color)

Question: What color is face 5 in the cube net shown in the image?

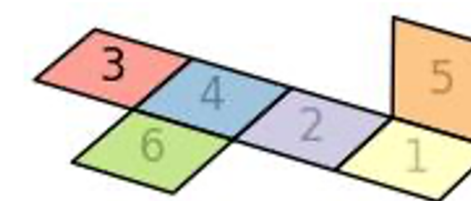
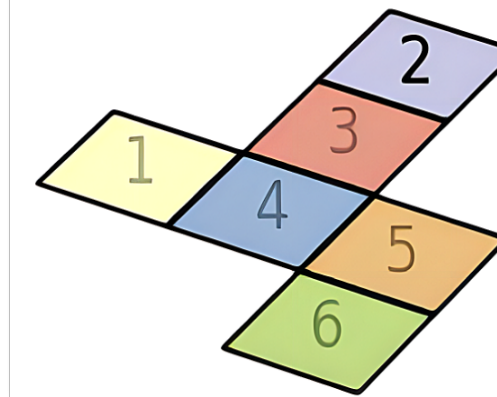
- A. Light orange
- B. Yellow
- C. Light purple
- D. Light red

2D Perception (Connectivity)

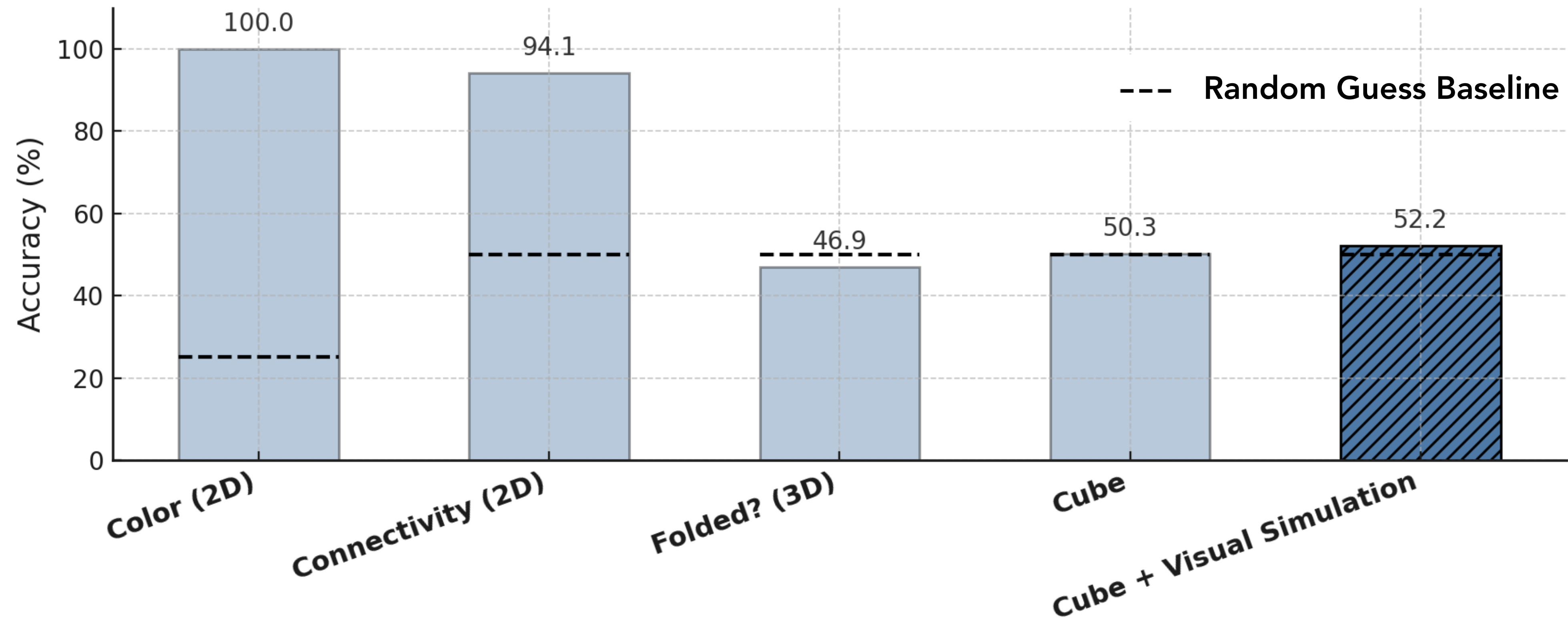
Question: In the cube net shown in the image, is face 2 directly connected to face 3, **yes** or no?

3D Perception (Folded?)

Question: In the cube net folding process shown below, has face 6 been folded towards 1 yet, **yes** or **no**?

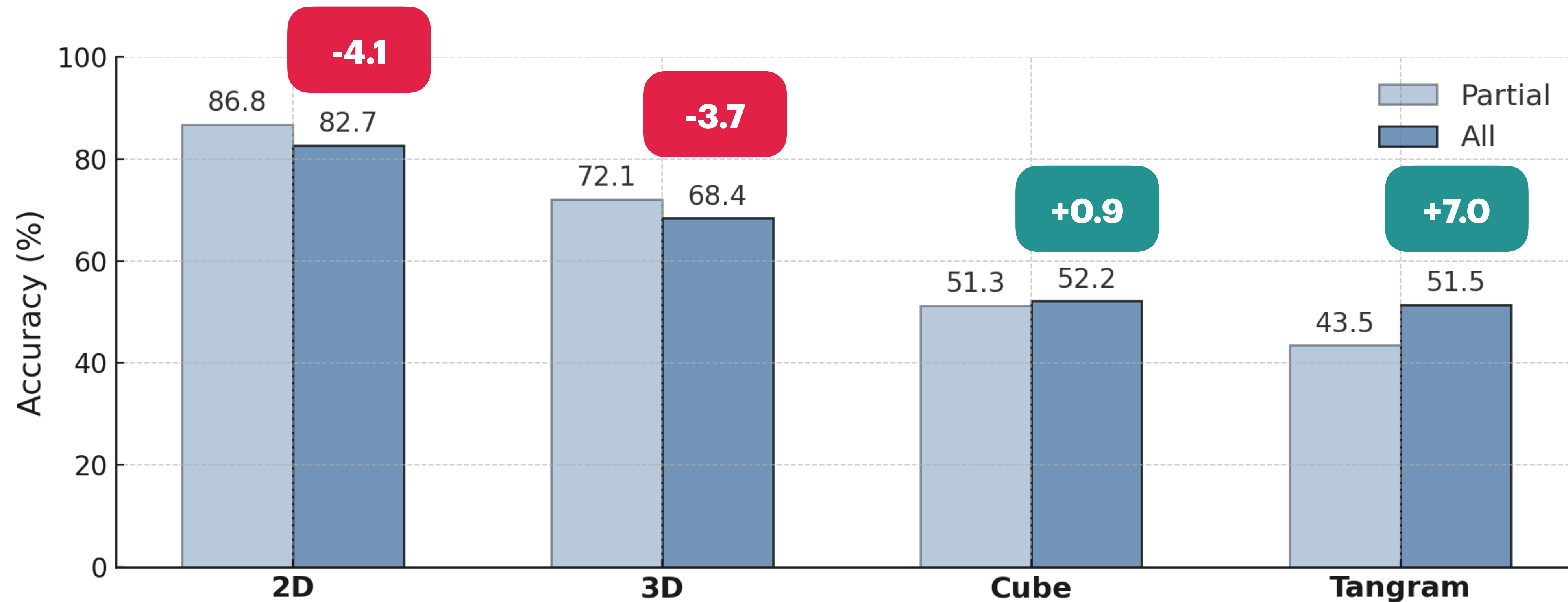


Models are Better at 2D Perception than 3D Perception



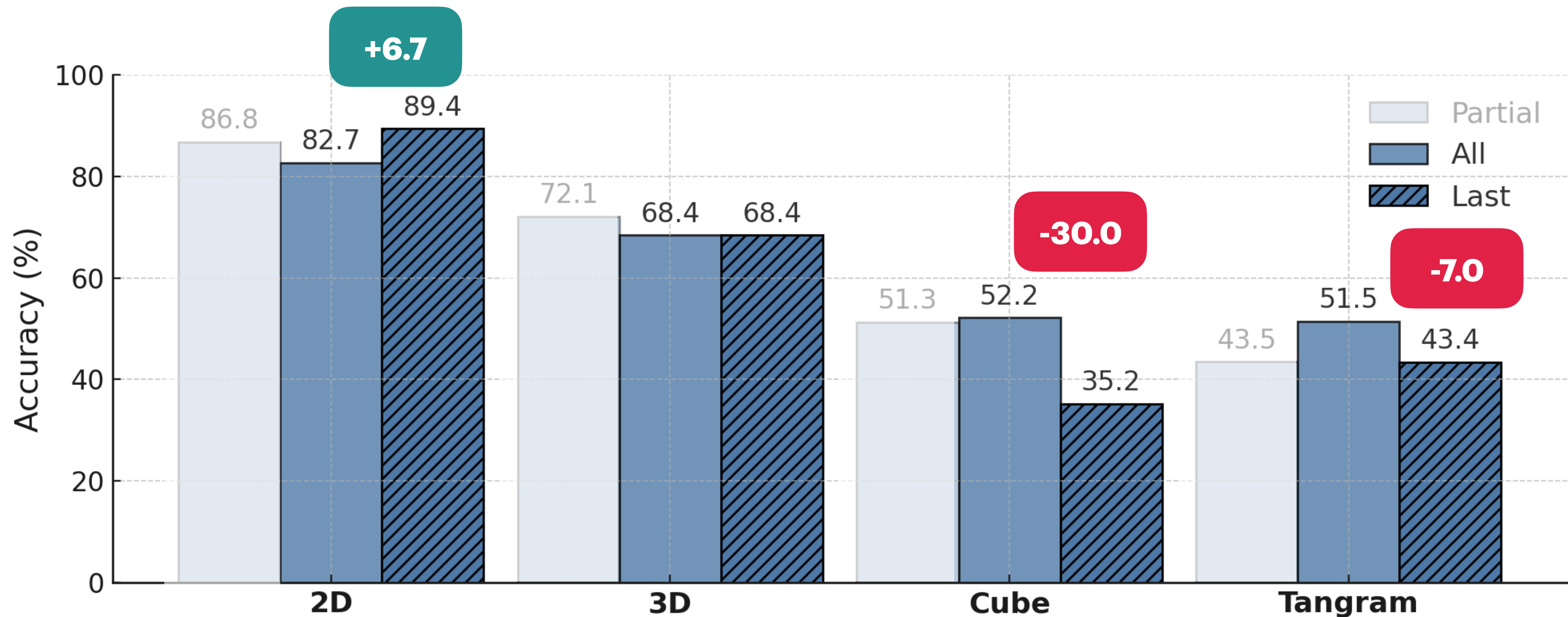
Do Models Need the Whole Visual Simulation Steps?

More challenging tasks benefit from the more complete visual simulations.



Do Models Need the Whole Visual Simulation Steps?

Last intermediate visual simulation only helps easier tasks.



Takeaways

We propose STARE to evaluate multimodal models on visual simulations

Visual simulation, though crucial to human reasoning, models today are not equipped with such capabilities

Models cannot effectively leverage explicit visual simulations for complex tasks, not to mention mentally simulating intermediate steps like how humans do

How do we teach models to perform visual simulation?
Explicit visual simulation or Implicit visual simulation